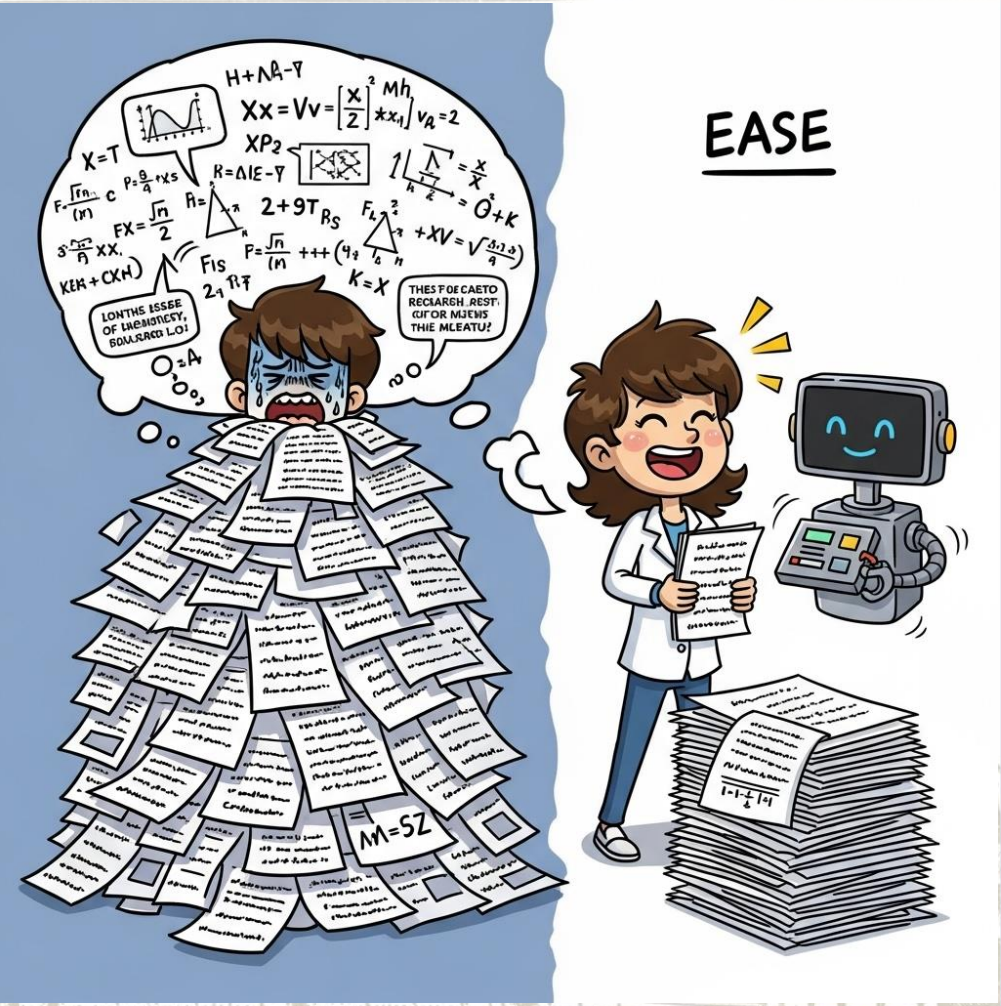
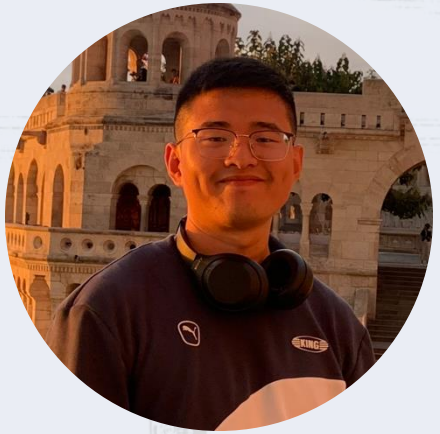
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



AI-Supported Literature Search

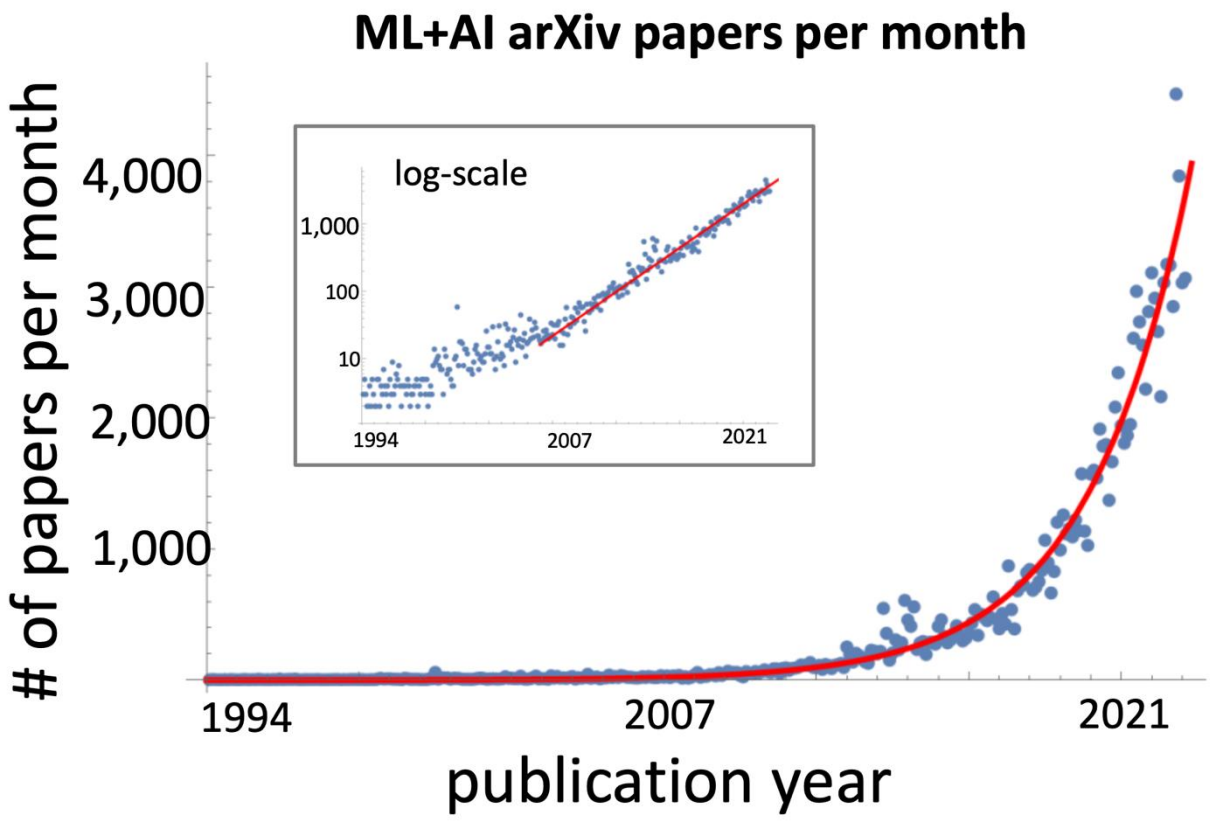
TRANSFORMING SCIENCE WITH LARGE LANGUAGE MODEL



Dr. Yong Cao
PostDoc, University of Tübingen
2025-07-26

* Photo from https://socialstudieshelp.com

Research Growth



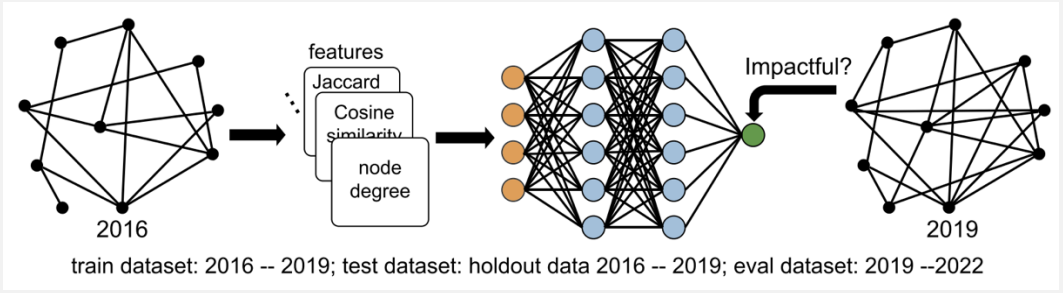
[Predicting the Future of AI with AI: High-Quality link prediction in an exponentially growing knowledge network, 2022]

*The real problem is **not information overload**, it's **filter failure**.*

— Clay Shirky

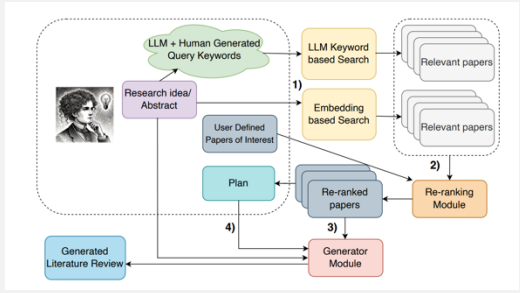
Why AI/LLMs for Literature Search?

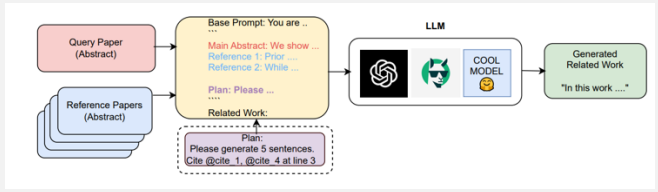
Uncover Emerging Trends



e.g., Forecasting high-impact research topics

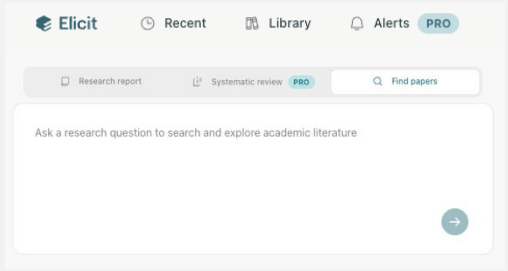
Semantic & Relation Understanding

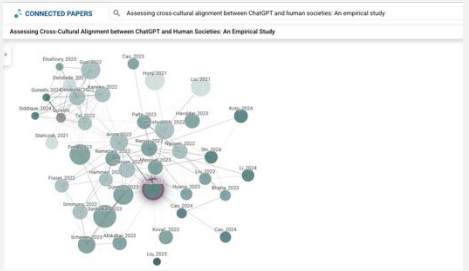


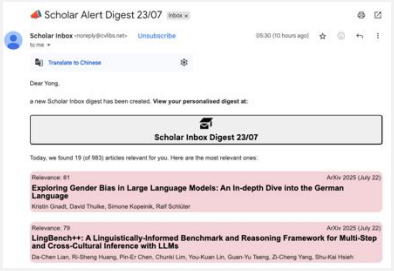


e.g., LLMs for literature review

Boost Research Efficiency and Increase Creativity



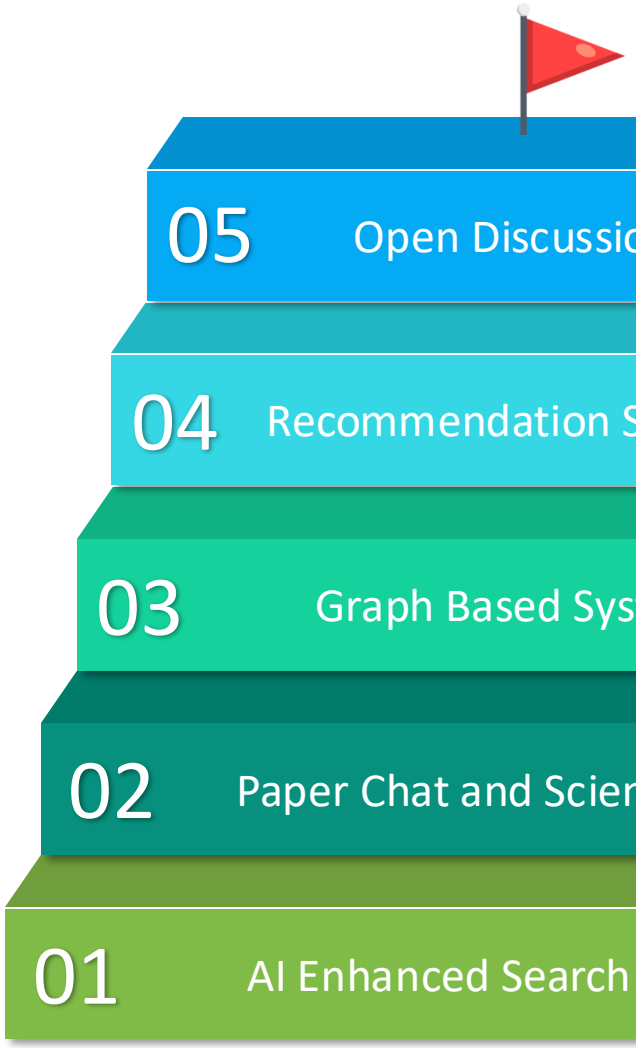




e.g., Scientific QA, Knowledge Graph Enhanced Search, Semantic Search, personalized Recommendation

Forecasting high-impact research topics via machine learning on evolving knowledge graphs, 2025
LitLLMs, LLMs for Literature Review: Are we there yet? 2025
Elicit, Connected Papers, Scholar Inbox, ResearchTrend.ai

Outline



05 Open Discussion — Future Direction and Conclusion

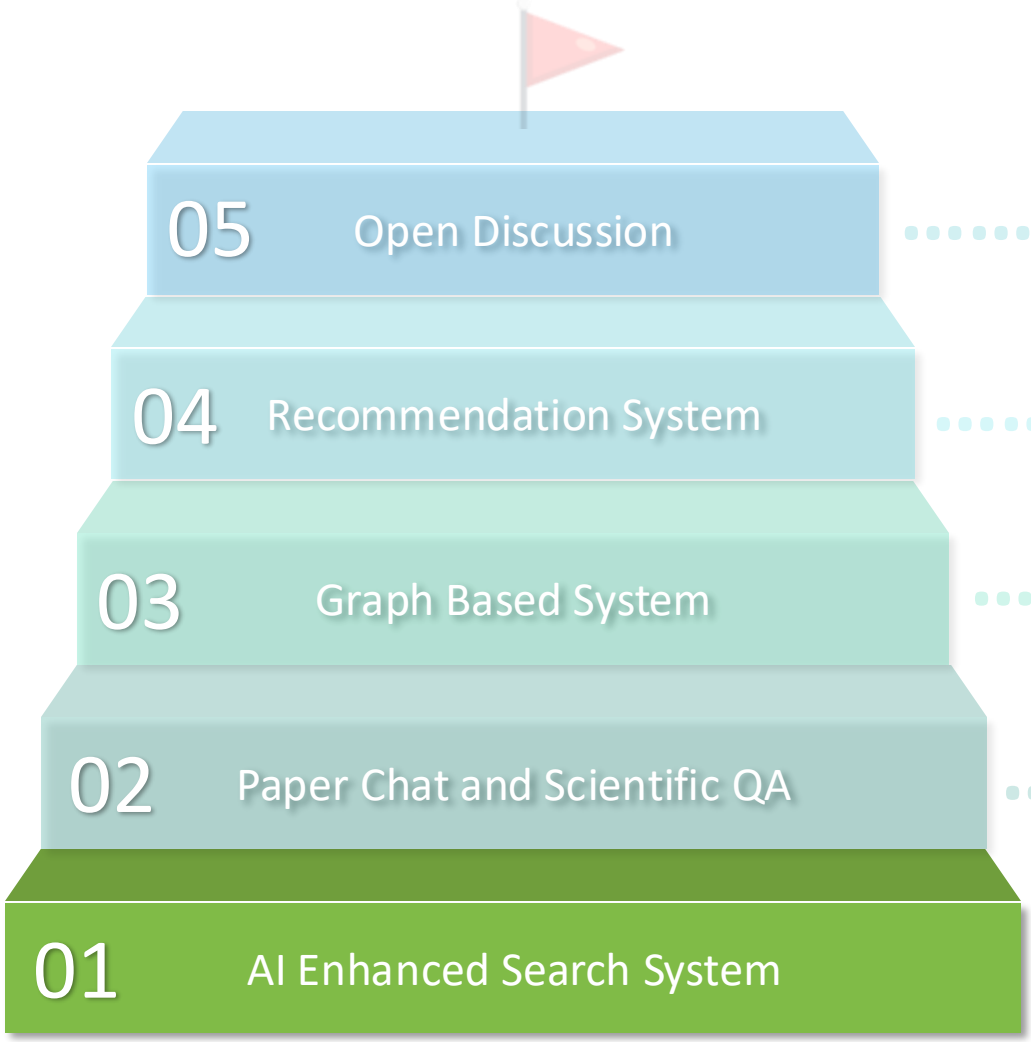
04 Recommendation System — Embedding, Content-Based Filtering, Collaborative Filtering

03 Graph Based System — Graph-based QA, SQL Search, Graph Learning, Visualization

02 Paper Chat and Scientific QA — RAG, Chunking, Multi-Agent Long Context Understanding

01 AI Enhanced Search System — Semantic Search, Information Extraction, Summarization

Outline



Future Direction and Conclusion

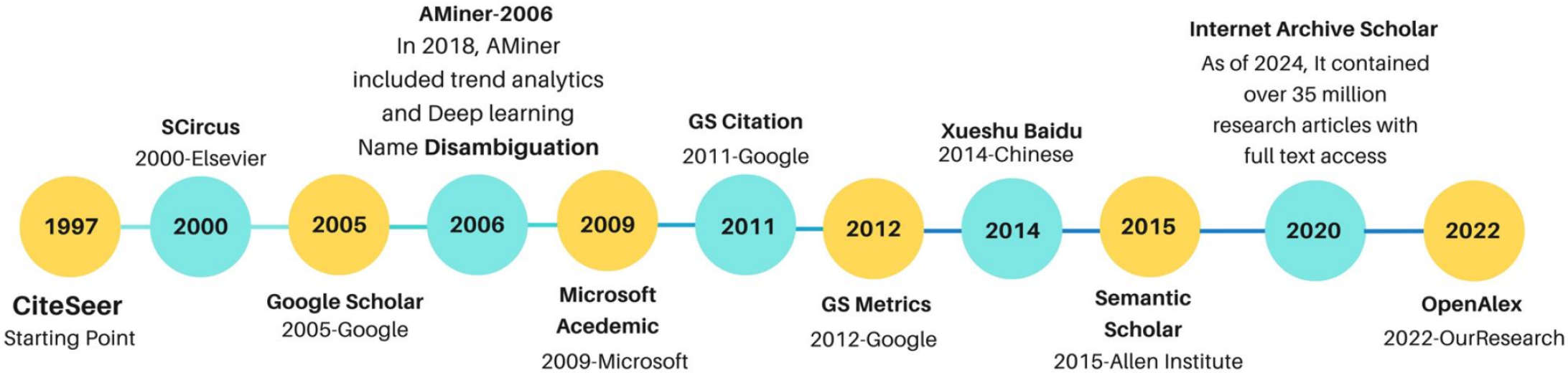
Embedding, Content-Based Filtering, Collaborative Filtering

Graph-based QA, SQL Search, Graph Learning, Visualization

RAG, Chunking, Multi-Agent Long Context Understanding

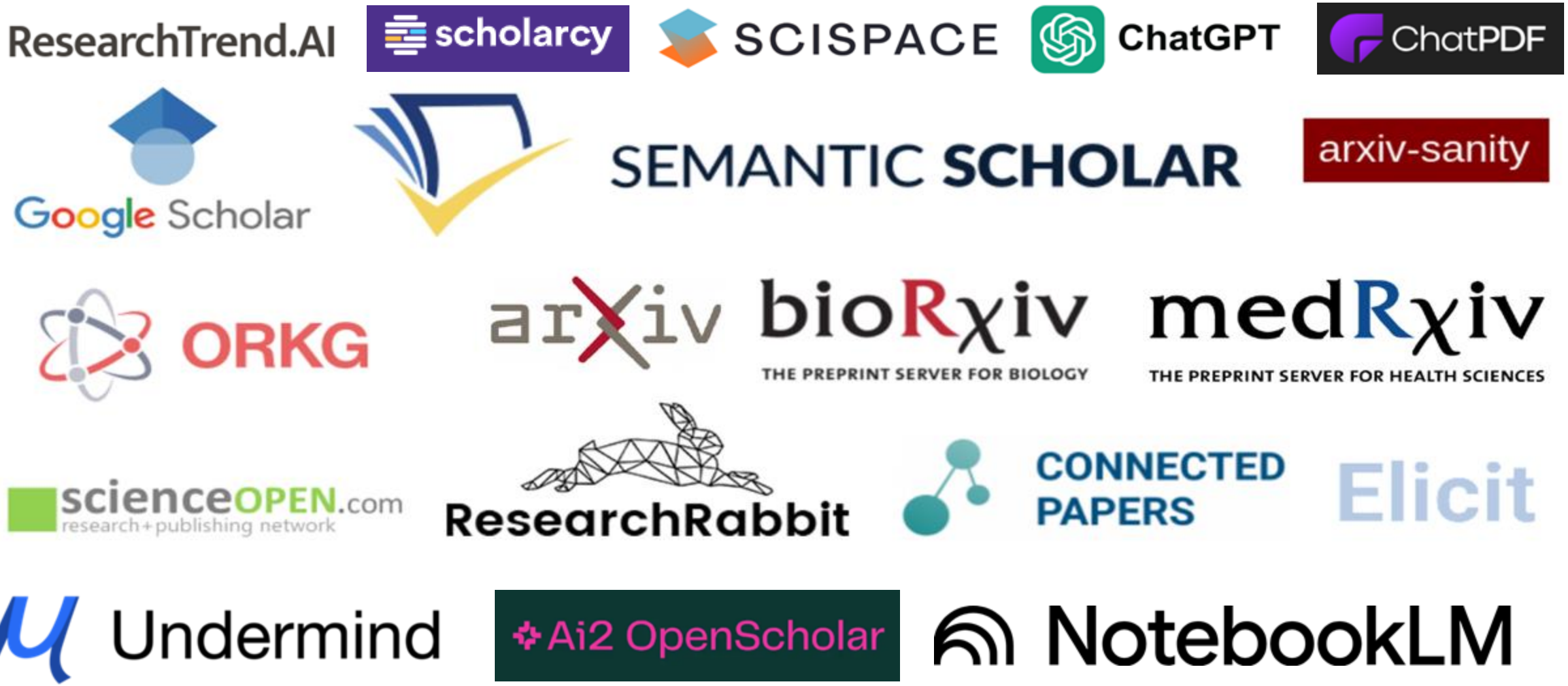
Semantic Search, Information Extraction, Summarization

Evolution of academic search engines

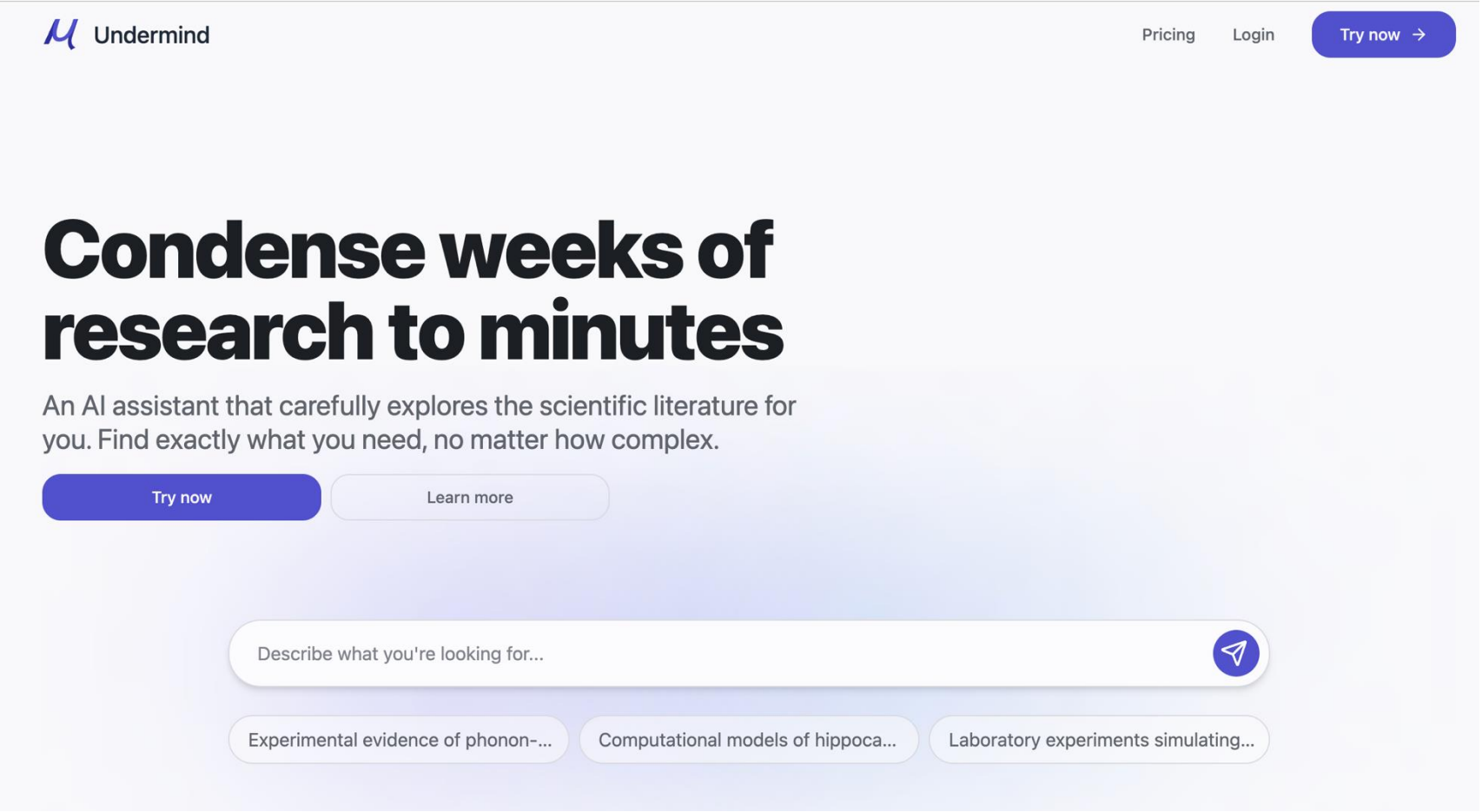


Comprehensive review of academic search systems: evolution, analysis, and future research directions, 2025

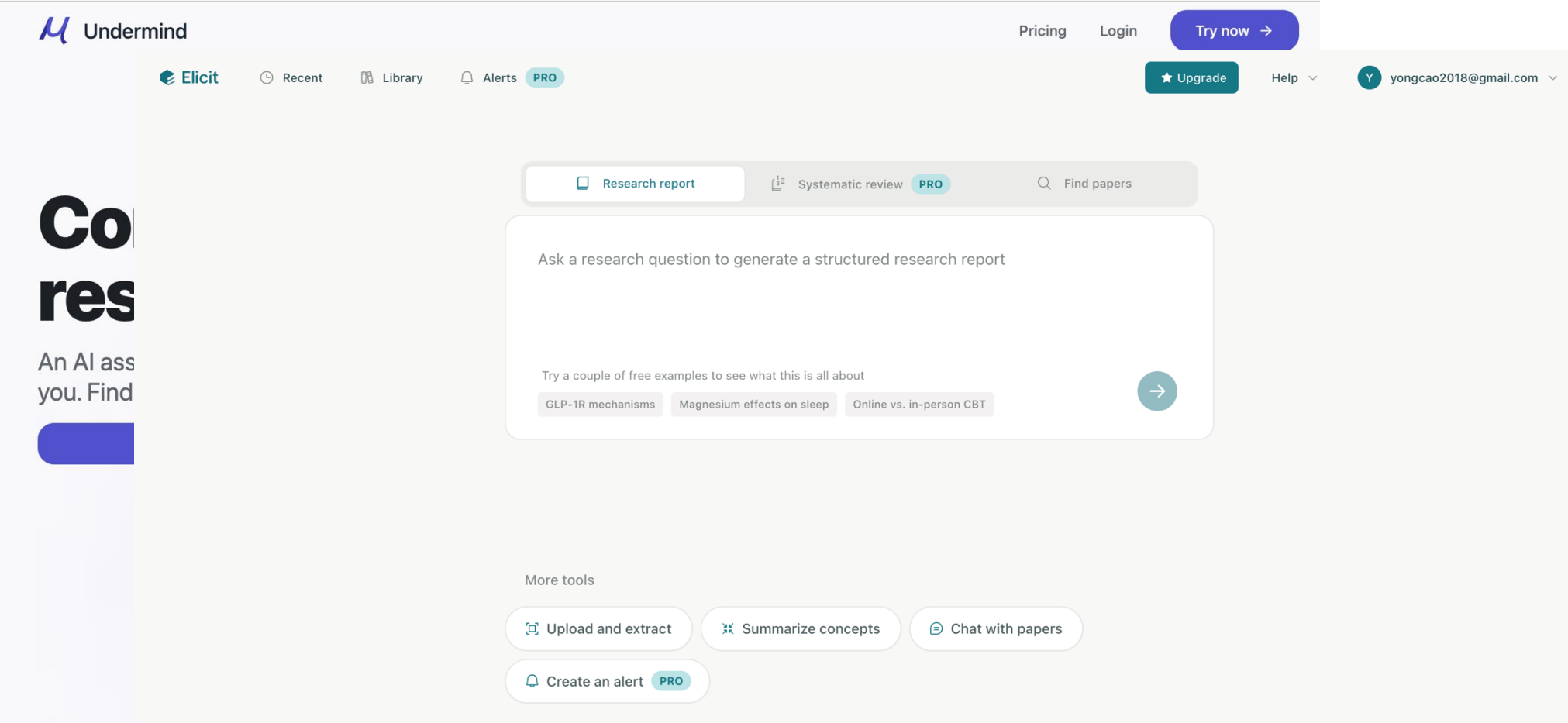
Data Source


ResearchTrend.AI
scholarcy
SCISPACE
ChatGPT
ChatPDF
Google Scholar
SEMANTIC SCHOLAR
arxiv-sanity
ORKG
arXiv
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
scienceOPEN.com
research+publishing network
ResearchRabbit
CONNECTED PAPERS
Elicit
Undermind
Ai2 OpenScholar
NotebookLM

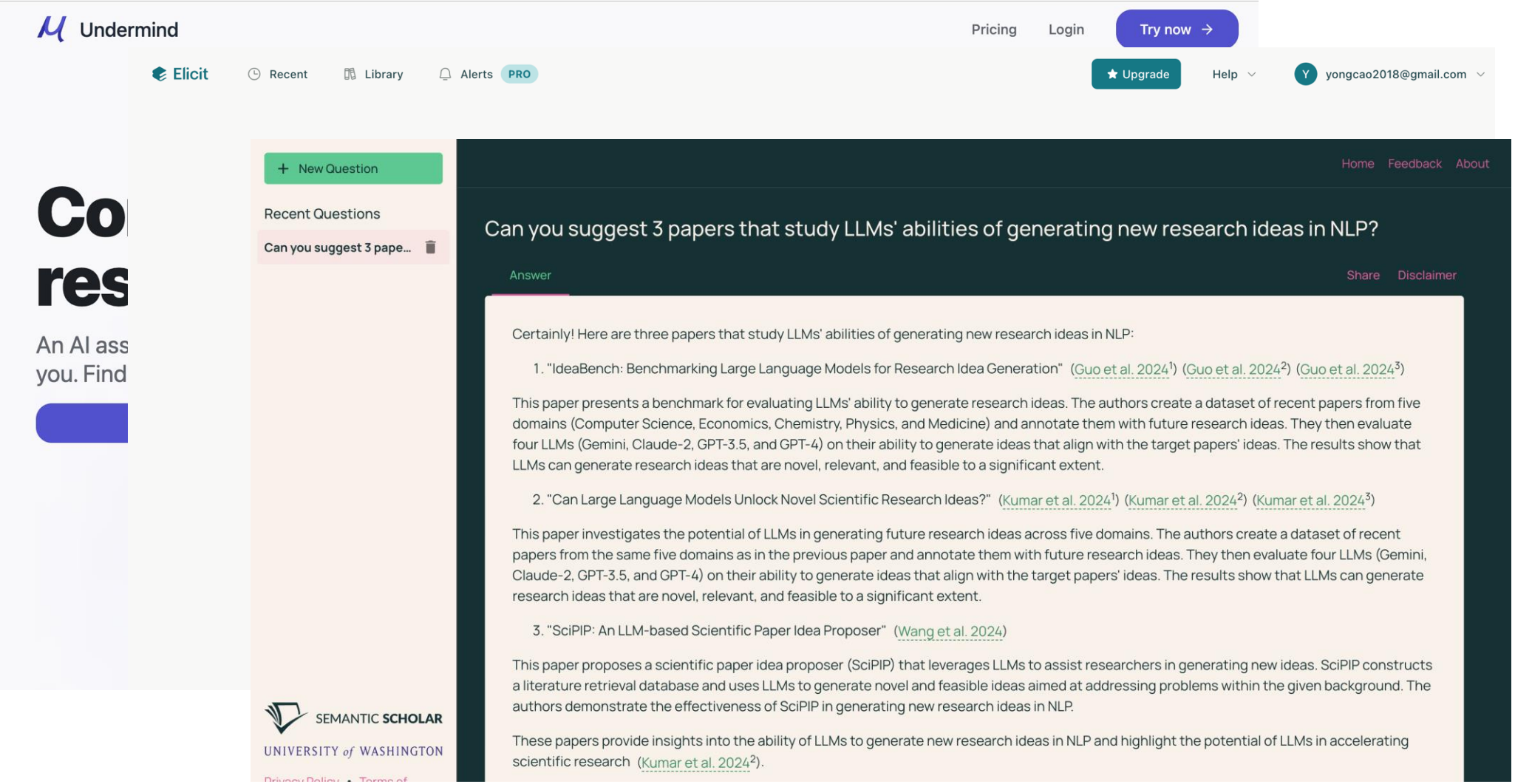
1 - AI Enhanced Search System



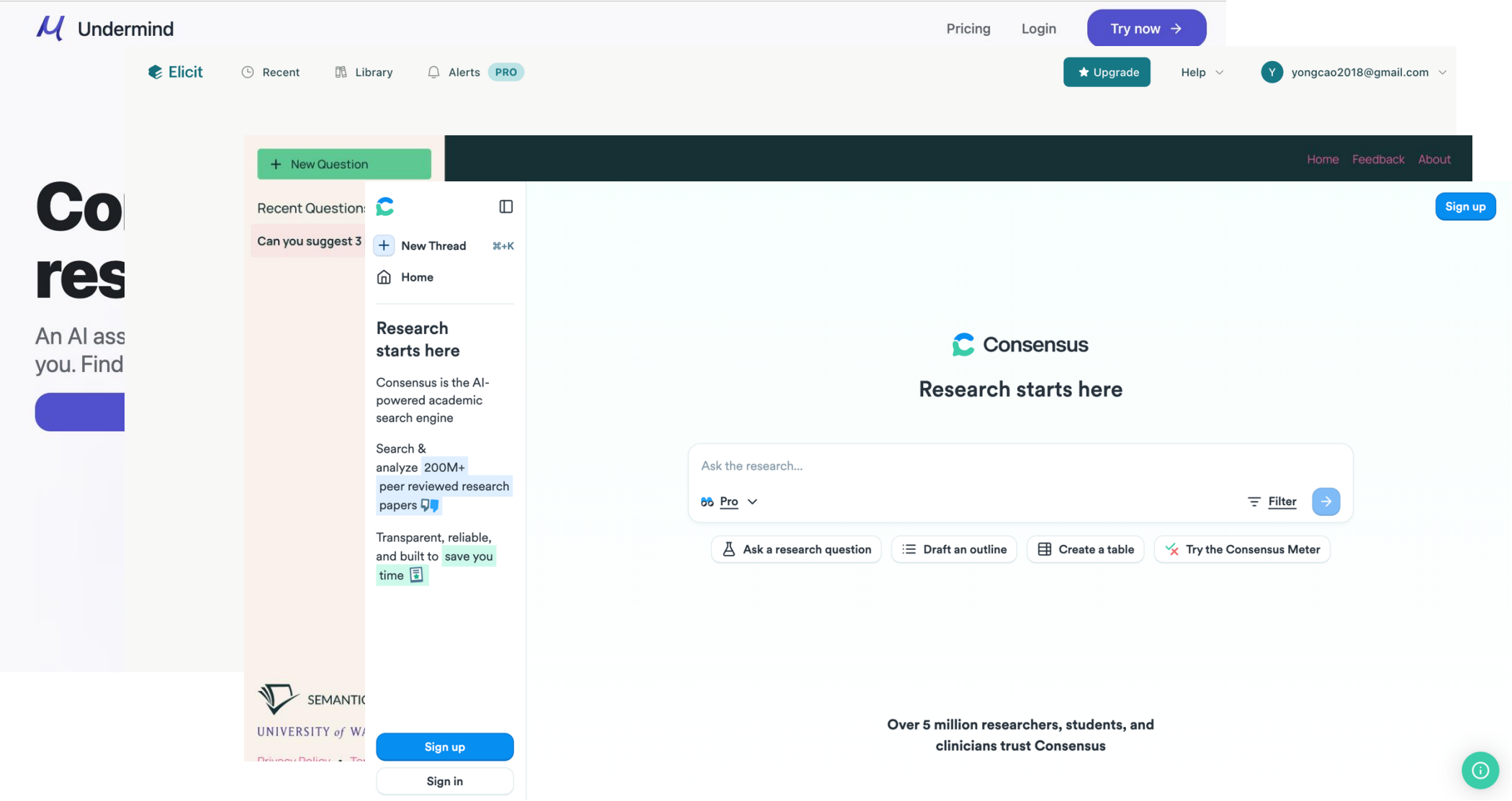
1 - AI Enhanced Search System



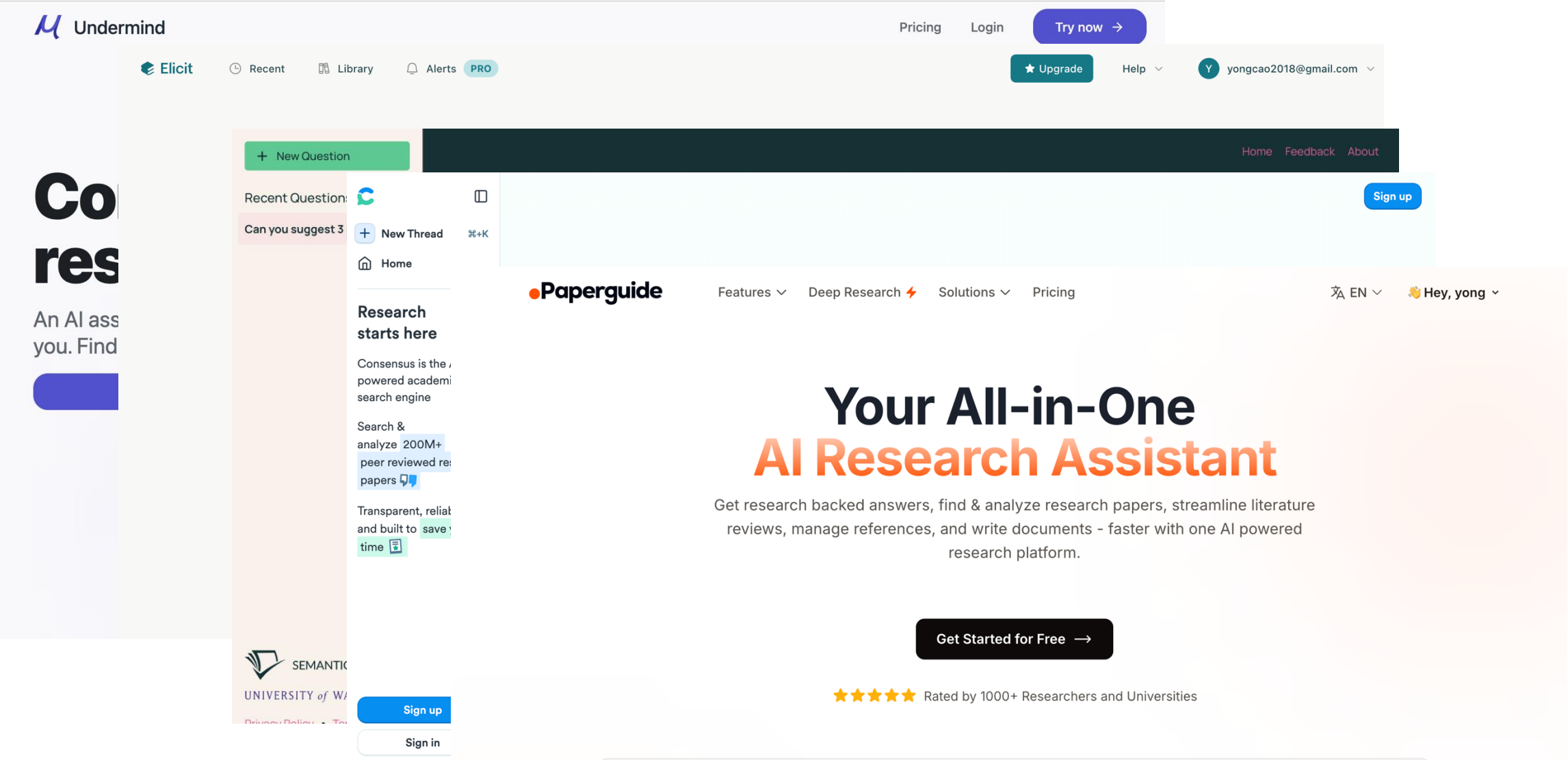
1 - AI Enhanced Search System



1 - AI Enhanced Search System



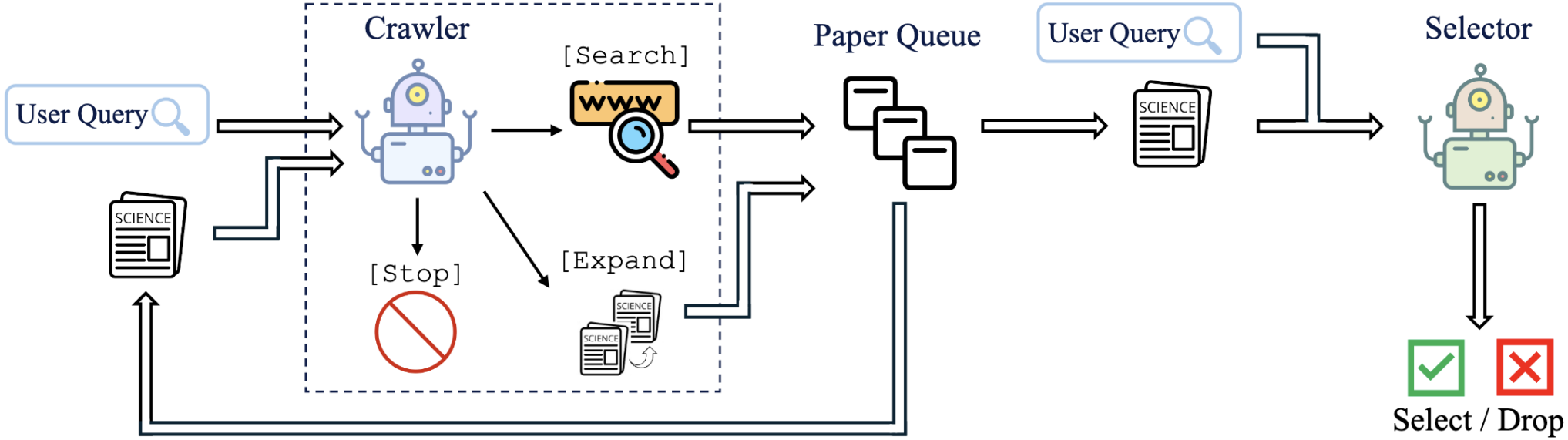
1 - AI Enhanced Search System



How AI enhance literature search platforms?

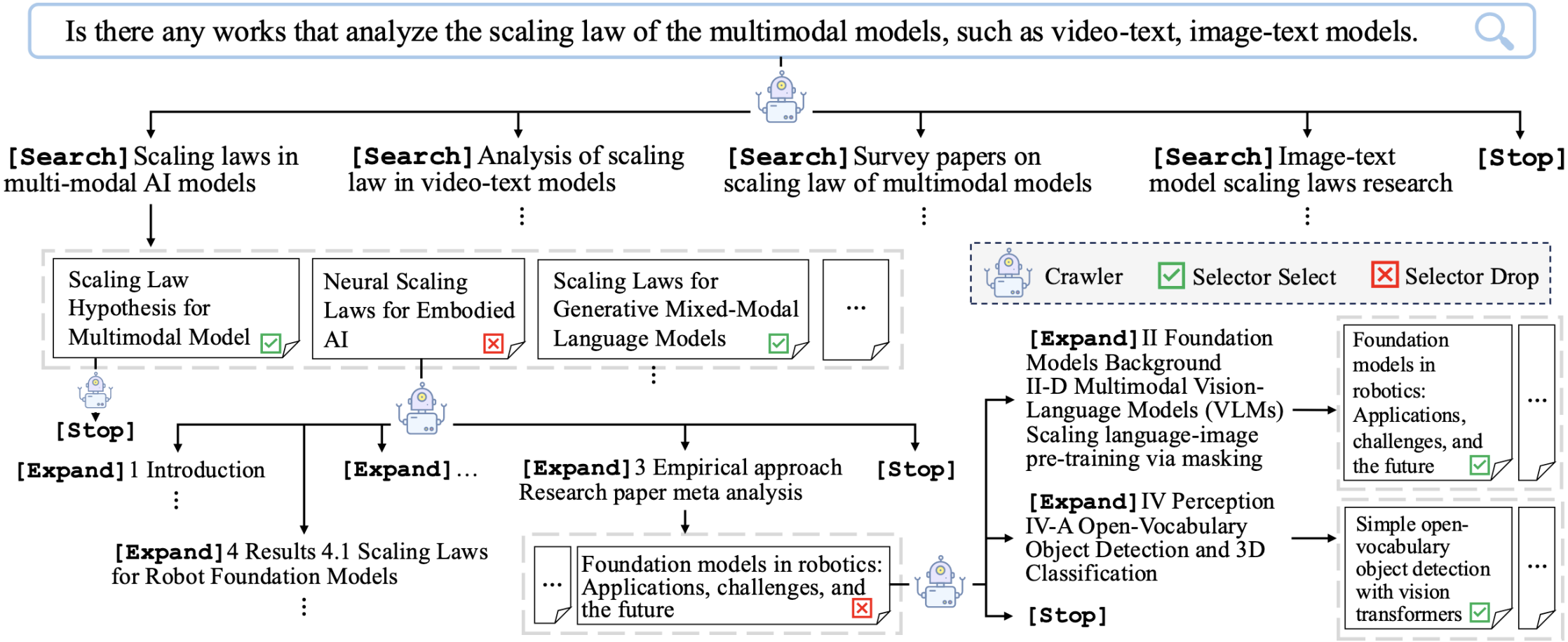
- Mimic human researchers workflows and expand search function.
- Two LLM agents: crawler and selector.

PaSa
Paper Search



PaSa: An LLM Agent for Comprehensive Academic Paper Search, 2025

How AI enhance literature search platforms?



Performance

- Imitation Learning + Reinforcement Learning

$$\mathcal{L}_{\text{policy}}(\theta) = \mathbb{E}_{\tau' \sim \pi_\theta^{\text{old}}} \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_\theta^{\text{old}}(a_t|s_t)} \hat{A}(s_t, a_t), \right. \right.$$
$$\left. \left. \text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_\theta^{\text{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}(s_t, a_t) \right) \right]$$

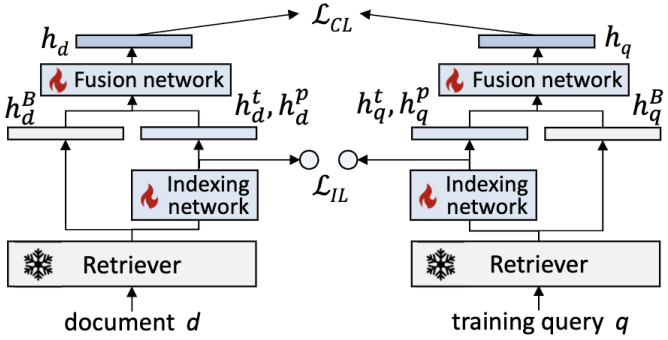
$$\mathcal{L}_{\text{value}}(\phi) = \mathbb{E}_{\tau' \sim \pi_\theta^{\text{old}}} \left[\max \left(\left(\hat{\mathrm{R}}_t - \hat{\mathrm{V}}_\phi(\mathrm{s}_t) \right)^2, \left(\hat{R}_t - \hat{V}_\phi^{\text{clip}}(s_t) \right)^2 \right) \right],$$

$$\mathcal{L}_{\text{RL}}(\theta, \phi) = \mathcal{L}_{\text{policy}}(\theta) + \eta \cdot \mathcal{L}_{\text{value}}(\phi)$$

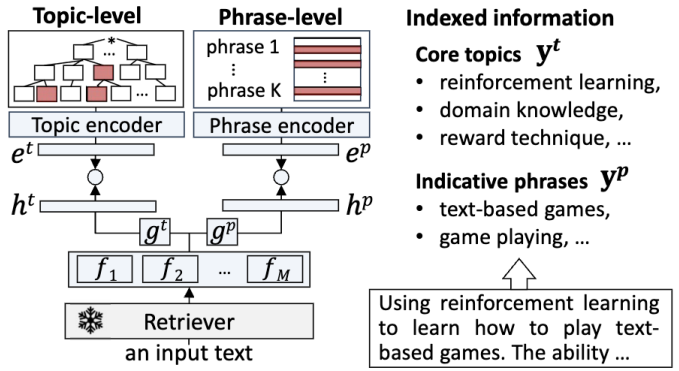
Method	Crawler Recall	Precision	Recall	Recall@100	Recall@50	Recall@20
Google	-	-	-	0.2015	0.1891	0.1568
Google Scholar	-	-	-	0.1130	0.0970	0.0609
Google with GPT-4o	-	-	-	0.2683	0.2450	0.1921
ChatGPT*	-	0.0507	0.3046	-	-	-
GPT-o1	-	0.0413	0.1925	-	-	-
PaSa-GPT-4o	0.7565	**0.1457**	0.3873	-	-	-
PaSa-7b	**0.7931**	**0.1448**	**0.4834**	**0.6947**	**0.6334**	**0.5301**
PaSa-7b-ensemble	**0.8265**	0.1410	**0.4985**	**0.7099**	**0.6386**	**0.5326**

Taxonomy-guided Index Construction

- The Problem: Beyond Surface-Level Text Matching
- TaxoIndex Framework
- Step 1: Constructing the Semantic Index
- Step 2: Index-grounded AI Training (Fine-tuning)
- Step 3: Enhanced Retrieval Process



(a) Index-grounded fine-tuning



(b) Index learning with the indexing network

		CSFCube						DORIS-MAE					
		N@5	N@10	M@5	M@10	R@50	R@100	N@5	N@10	M@5	M@10	R@50	R@100
	BM25	0.307	0.310	0.088	0.134	0.504	0.635	0.354	0.330	0.079	0.107	0.490	0.669
SPECTER-v2	no Fine-Tuning	0.352	0.337	0.108	0.151	0.524	0.680	0.385	0.360	0.079	0.113	0.551	0.709
	FFT	0.372	0.368	0.123	0.169	0.576	0.692	0.408	0.387	0.084	0.122	0.562	0.736
	aFT	0.378	0.344	0.119	0.160	0.578	0.696	0.400	0.372	0.080	0.115	0.558	0.714
	FFT w/ GRF	0.331	0.317	0.112	0.152	0.561	0.705	0.400	0.379	0.087	0.123	**0.586**	**0.756**
	FFT w/ ToTER	0.406	0.375	0.135	0.179	0.591	0.710	0.423	0.394	0.091	0.128	0.563	0.736
	JTR	0.379	0.352	0.118	0.157	0.598	0.699	0.395	0.380	0.080	0.118	0.548	0.713
	TaxoIndex	0.458†*	0.417†*	0.144†*	0.198†*	**0.633**†*	0.741†*	0.447†*	0.421†*	0.104†*	0.144†*	0.578†	**0.756**†
	TaxoIndex ++	**0.469**†*	**0.426**†*	**0.158**†*	**0.209**†*	0.621†*	**0.746**†*	**0.449**†*	**0.424**†*	**0.105**†*	**0.145**†*	0.581†	0.751†

Taxonomy-guided Semantic Indexing for Academic Paper Search, 2024

Deep Research -- ChatGPT / Gemini

Model data and training

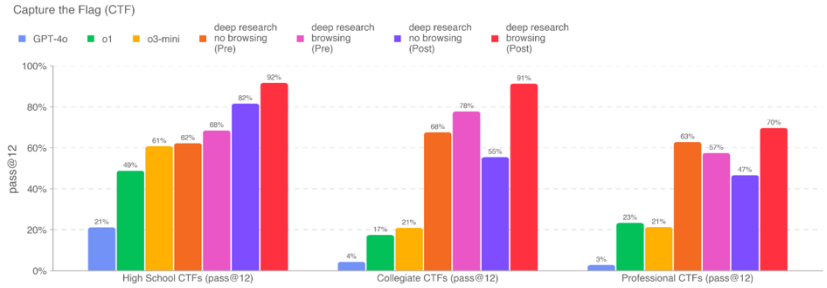
- Browsing datasets
- Graded against the ground truth answers or chain-of-thought model
- Safety datasets from o1 training

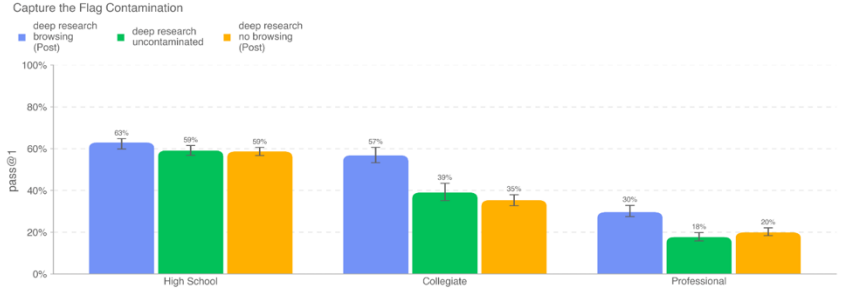
Risk and mitigation

- Prompt Injections
- Disallowed Content
- Privacy
- Ability to Run Code
- Bias
- Hallucinations

Overall Evaluation: Medium

OpenAI





Planning

Deep Research transforms your prompt into a personalized multi-point research plan

Gemini

Searching

Deep Research autonomously searches and deeply browses the web to find relevant, up-to-date information

Reasoning

Deep Research shows its thoughts as it reasons over information gathered iteratively and thinks before making its next move

Reporting

Deep Research provides comprehensive custom research reports with more detail and insights, generated in minutes and available as an Audio Overview, saving you hours of time

Deep Research System Card, 2025
Gemini Deep Research

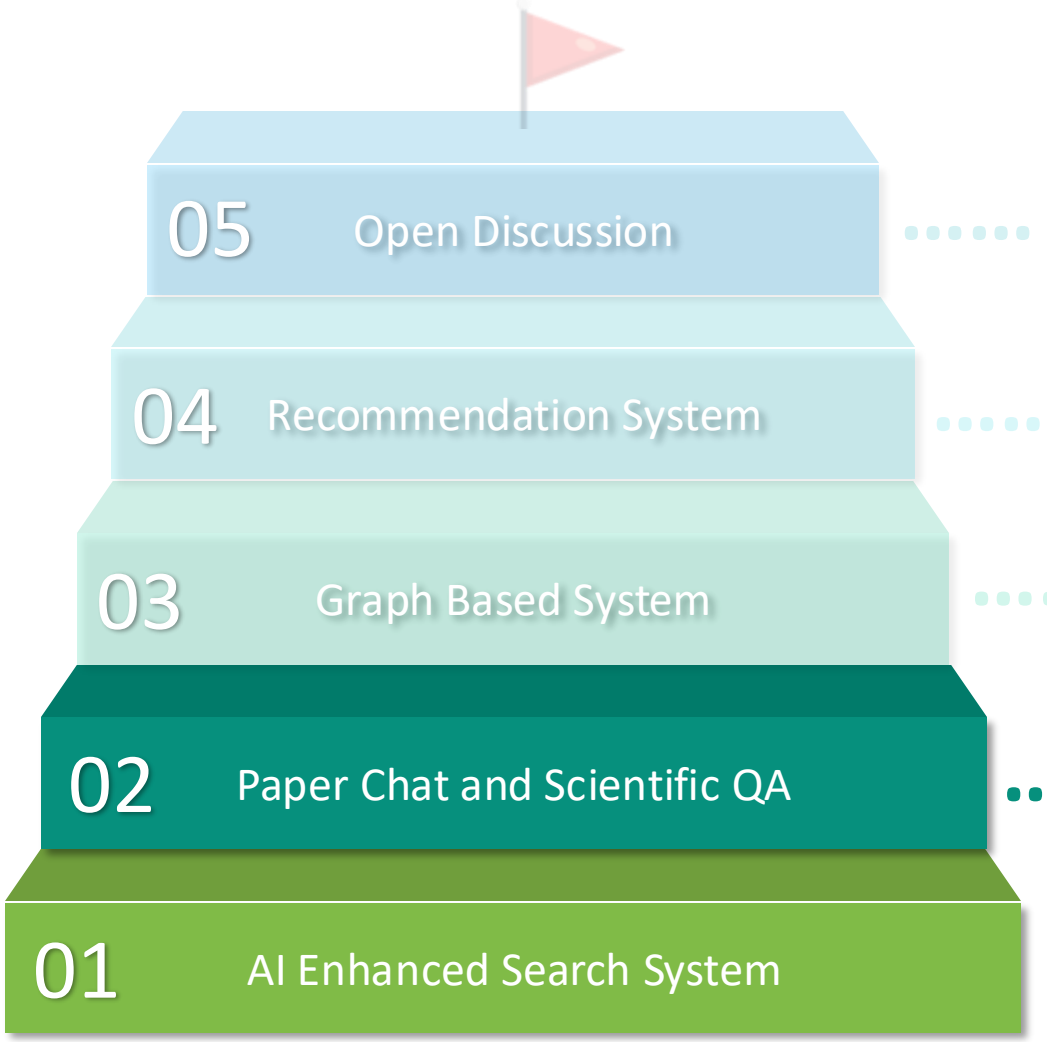
1 (Summary) - AI Enhanced Search System

- Main Function
  - Optimize scholarly information retrieval
  - Context-aware, semantically rich, and personalized search results

- Key Techniques
  - LLM-based agents
  - Embedding-based Retrieval
  - Personalization

- Challenges
  - Data heterogeneity
  - Limited handling of complex scientific content

Outline



Future Direction and Conclusion

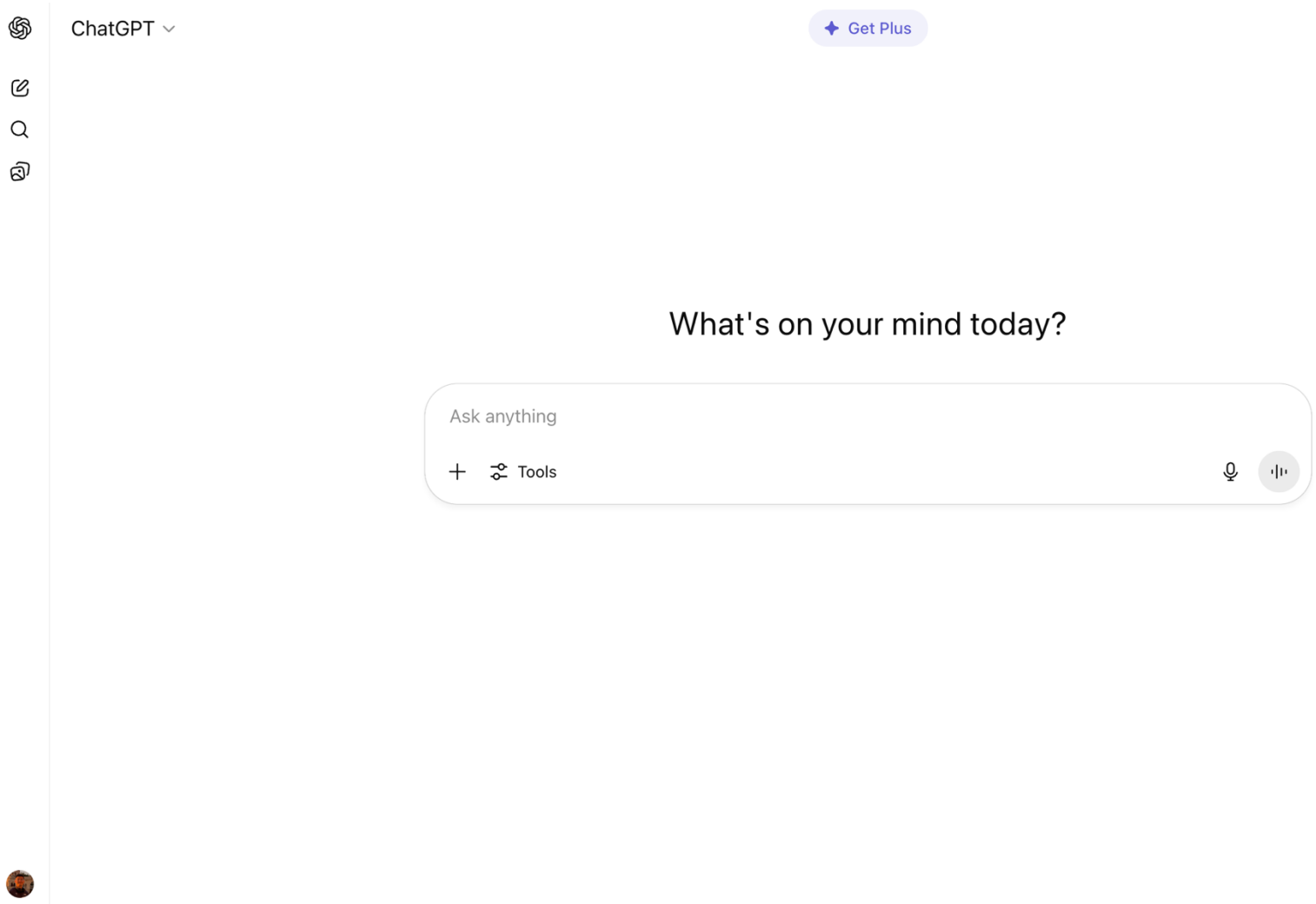
Embedding, Content-Based Filtering, Collaborative Filtering

Graph-based QA, SQL Search, Graph Learning, Visualization

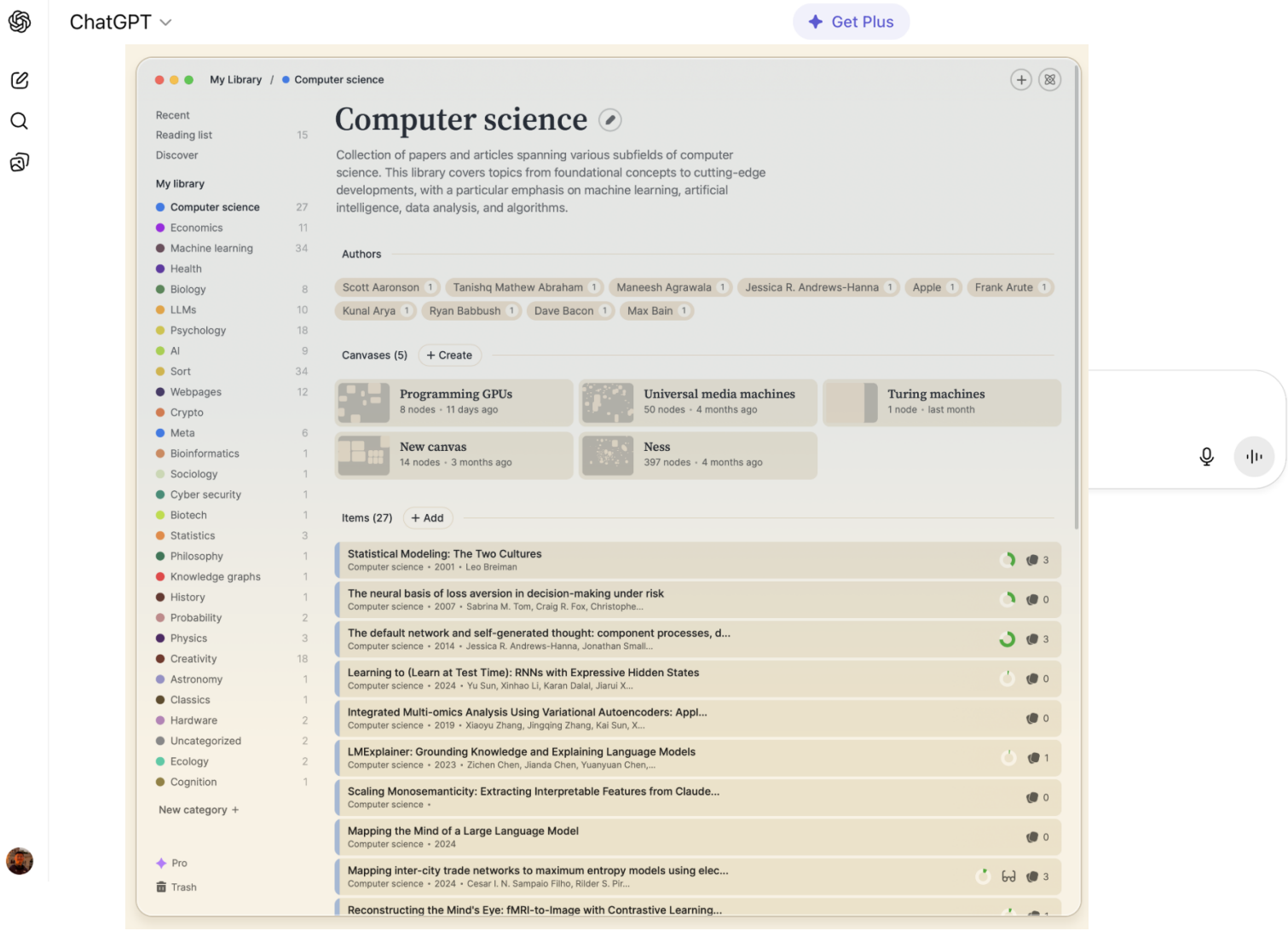
RAG, Chunking, Multi-Agent Long Context Understanding

Semantic Search, Information Extraction, Summarization

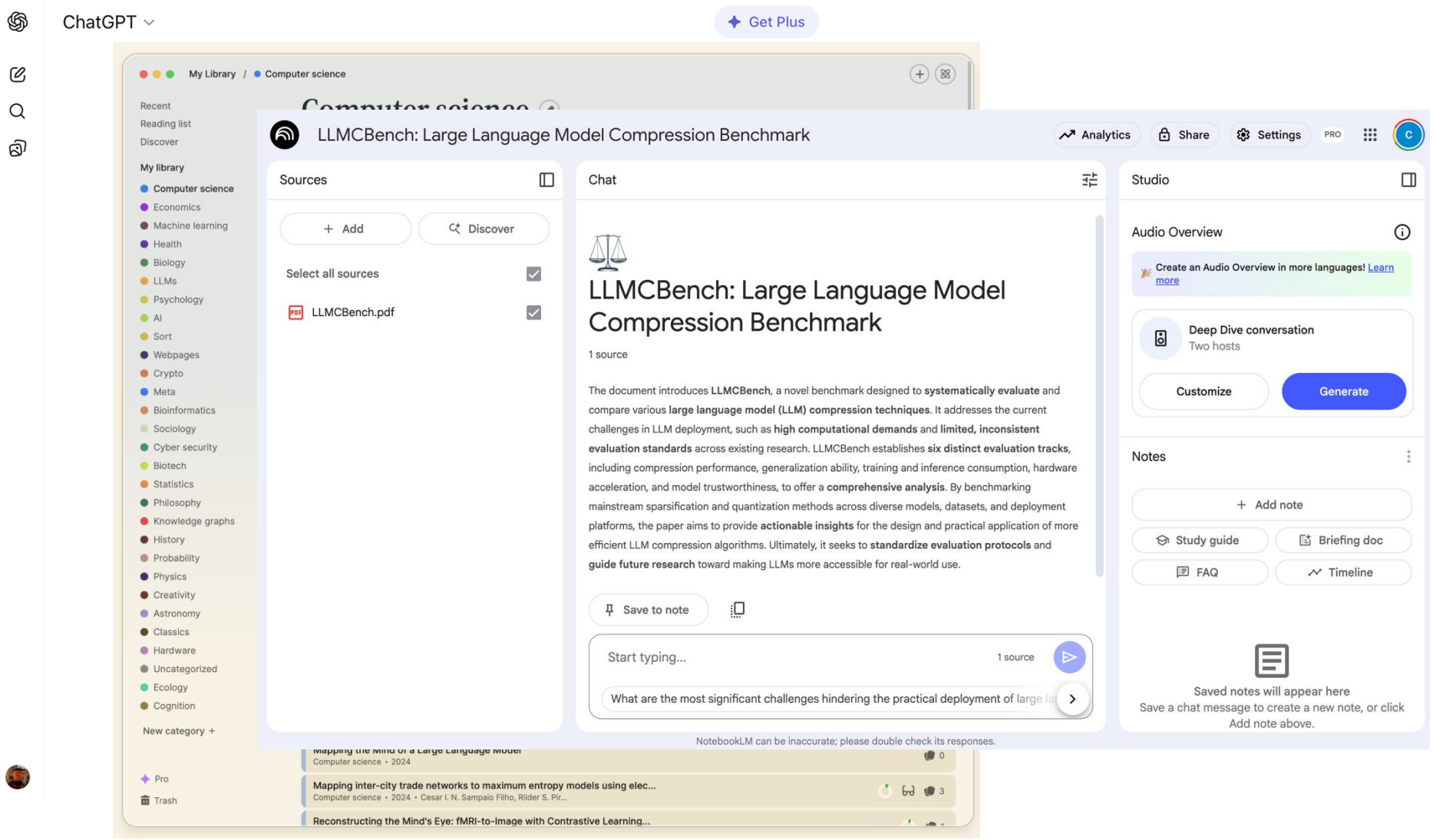
2 - Paper Chat and Scientific QA



2 - Paper Chat and Scientific QA



2 - Paper Chat and Scientific QA

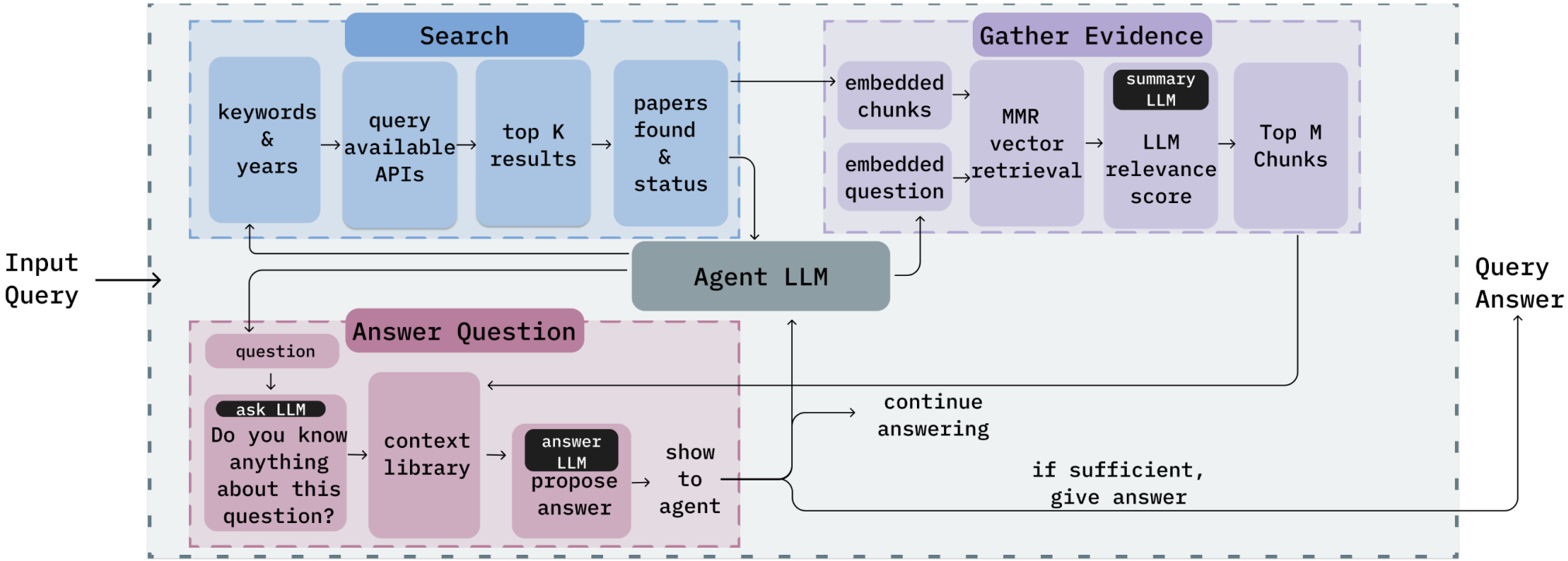


2 - Paper Chat and Scientific QA

2 - Paper Chat and Scientific QA

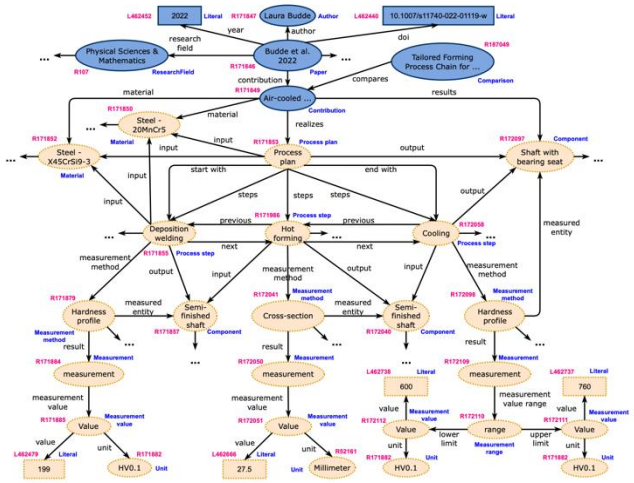
PaperQA – An Agent-Based RAG System

- Aims to address the hallucinations and knowledge update.
- Key Components:
  - Search, Gather Evidence, Answer Question

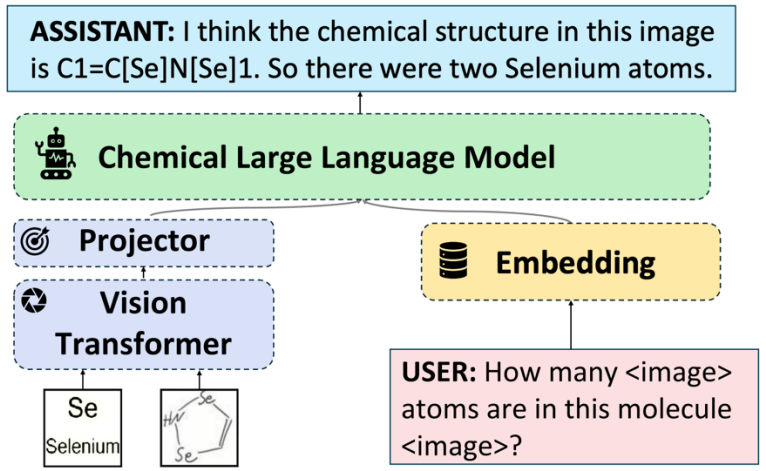


PaperQA: Retrieval-Augmented Generative Agent for Scientific Research, 2023

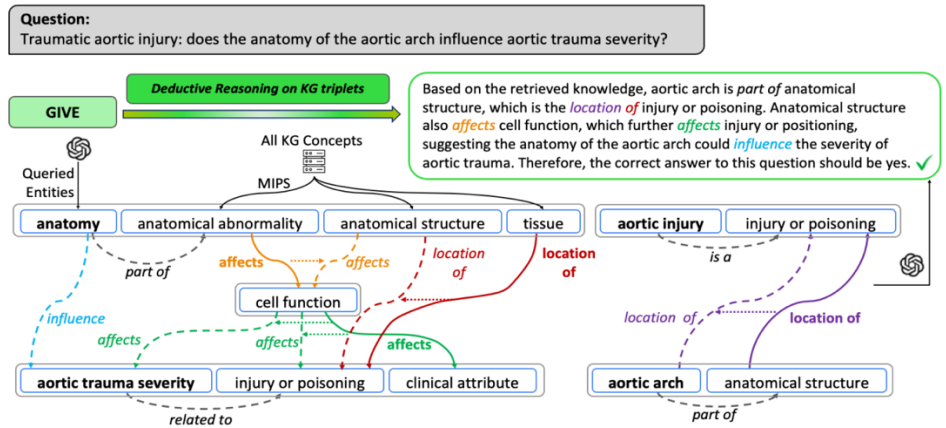
Mostly focus on…



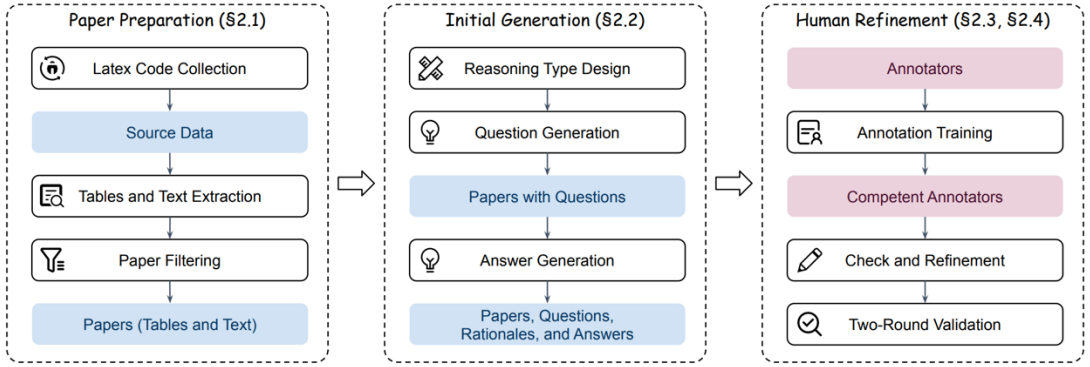
KG Integration



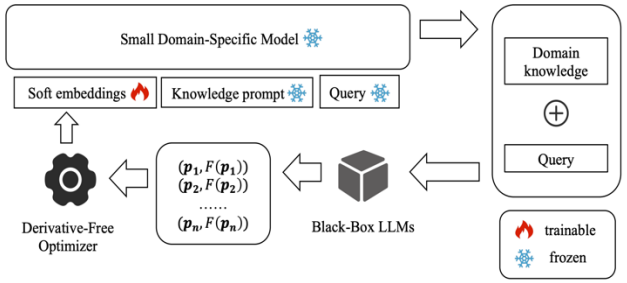
Multi-domal QA



Complex Reasoning



Benchmarking



Domain knowledge

[The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge, 2023]
[ChemVLM: Exploring the Power of Multimodal Large Language Models in Chemistry Area, 2025]
[BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models, 2024]
[SCITAT: A Question Answering Benchmark for Scientific Tables and Text Covering Diverse Reasoning Types, 2024]
[GIVE: Structured Reasoning of Large Language Models with Knowledge-Graph-Inspired Veracity Extrapolation, 2025]

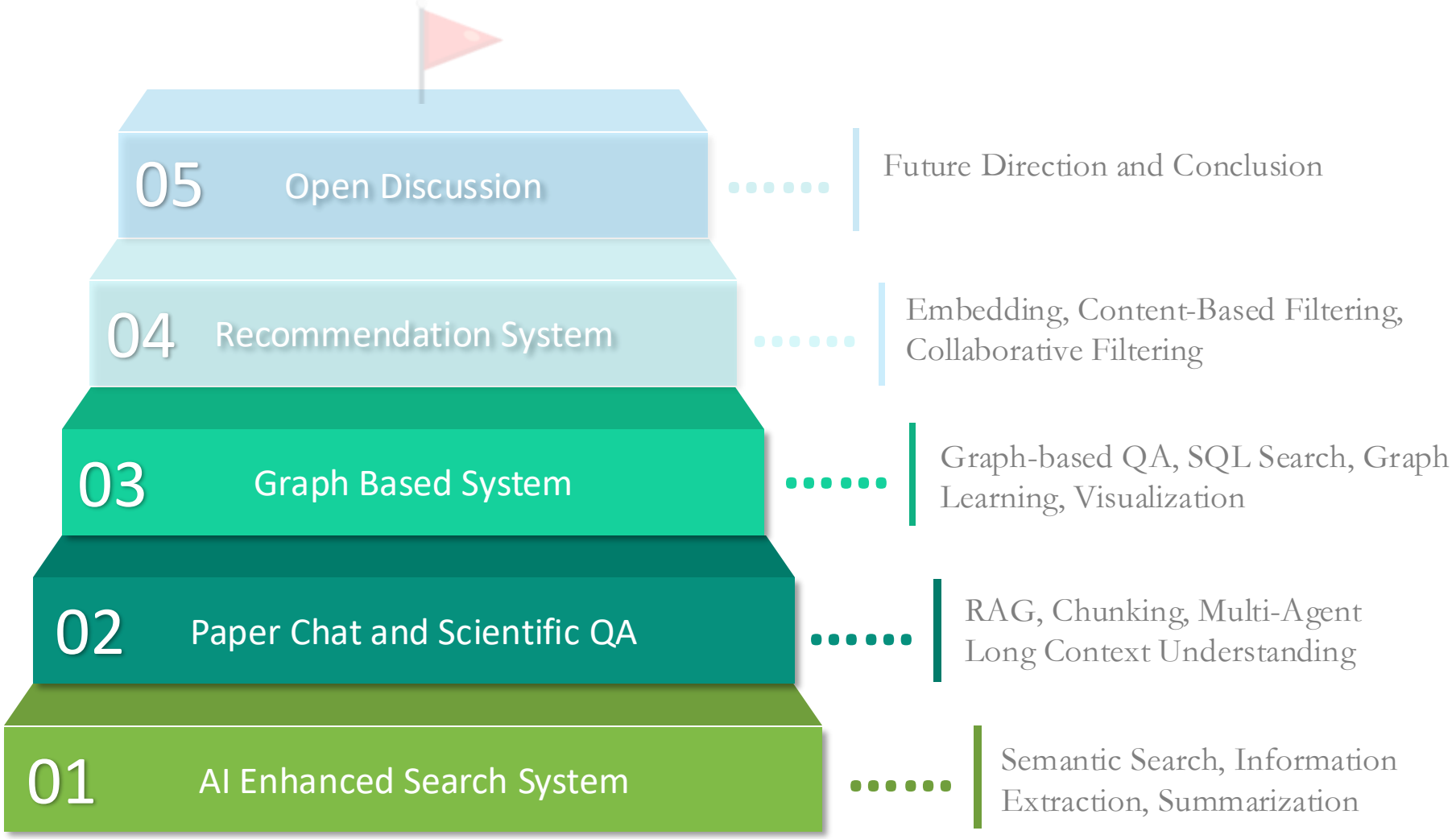
2 (Summary) - Paper Chat and Scientific QA

- Main Function
  - Interactive Q&A with research papers

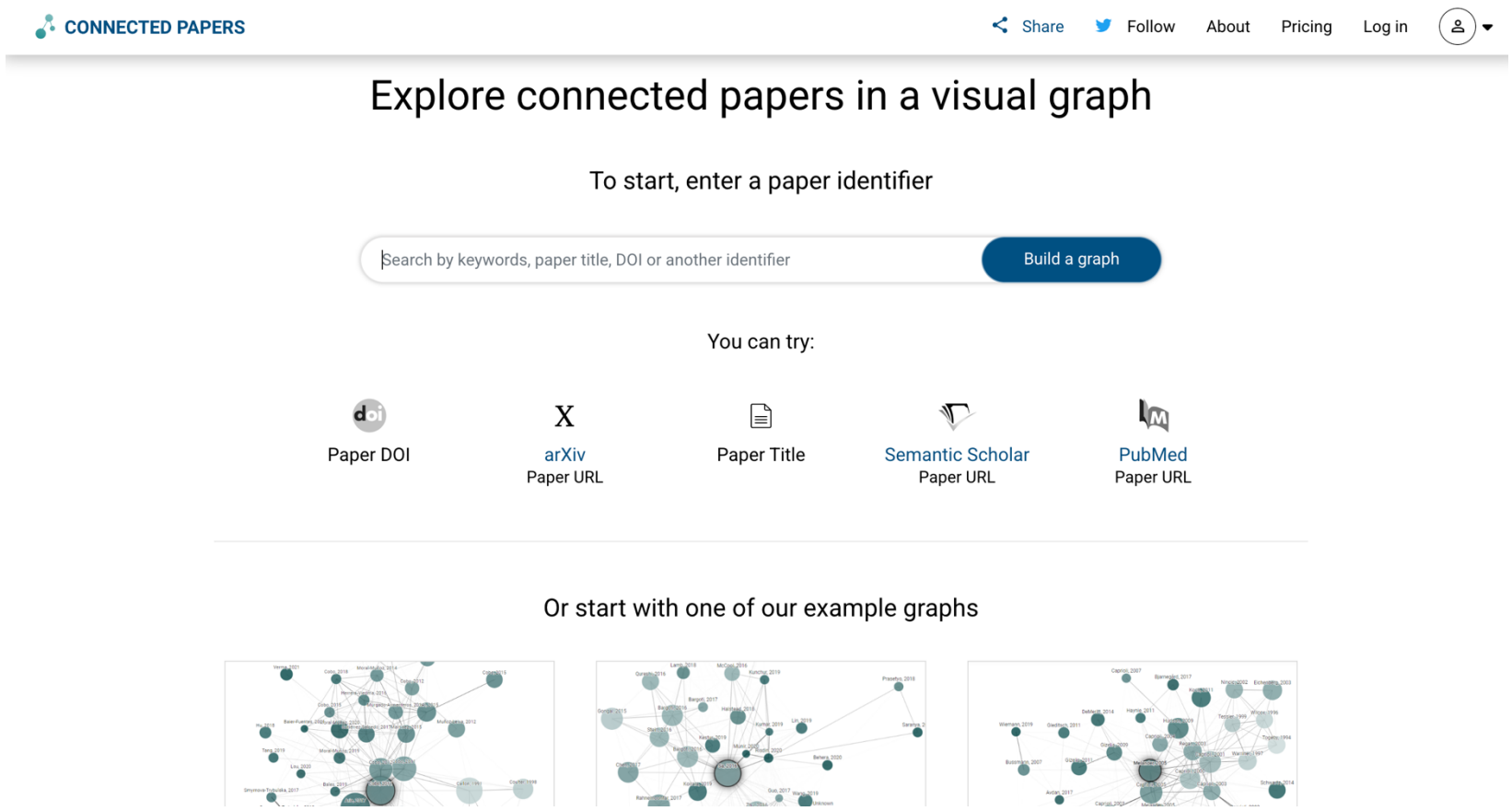
- Key techniques
  - LLM-based agents
  - PDF Parsing & Contextual Chunking
  - Summarization & Key Point Extraction
  - Multi-turn Dialogue Tracking

- Challenges
  - PDF Quantity & Size Limits
  - Domain Knowledge
  - Reliability & Explainability

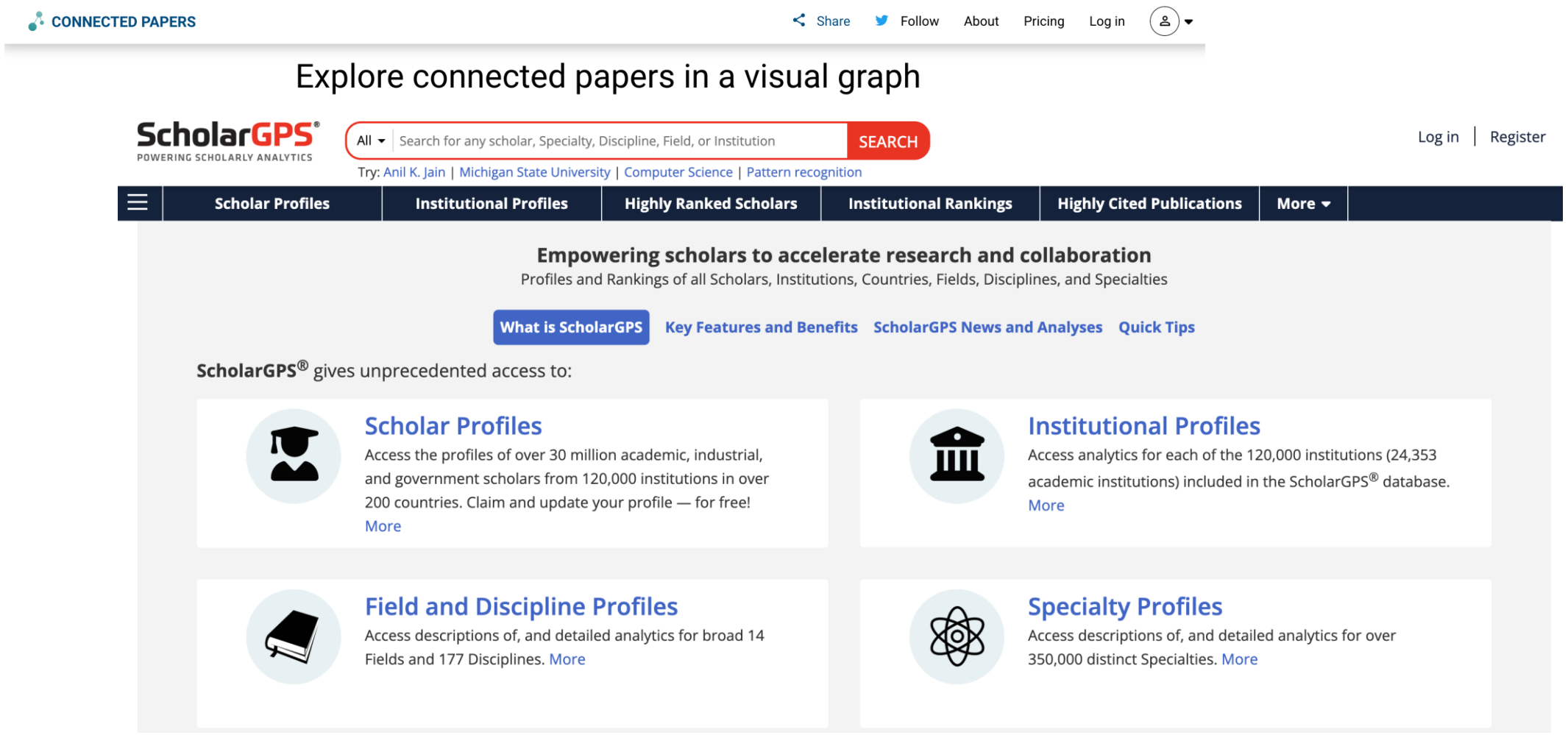
Outline


05
Open Discussion
Future Direction and Conclusion
04
Recommendation System
Embedding, Content-Based Filtering, Collaborative Filtering
03
Graph Based System
Graph-based QA, SQL Search, Graph Learning, Visualization
02
Paper Chat and Scientific QA
RAG, Chunking, Multi-Agent Long Context Understanding
01
AI Enhanced Search System
Semantic Search, Information Extraction, Summarization

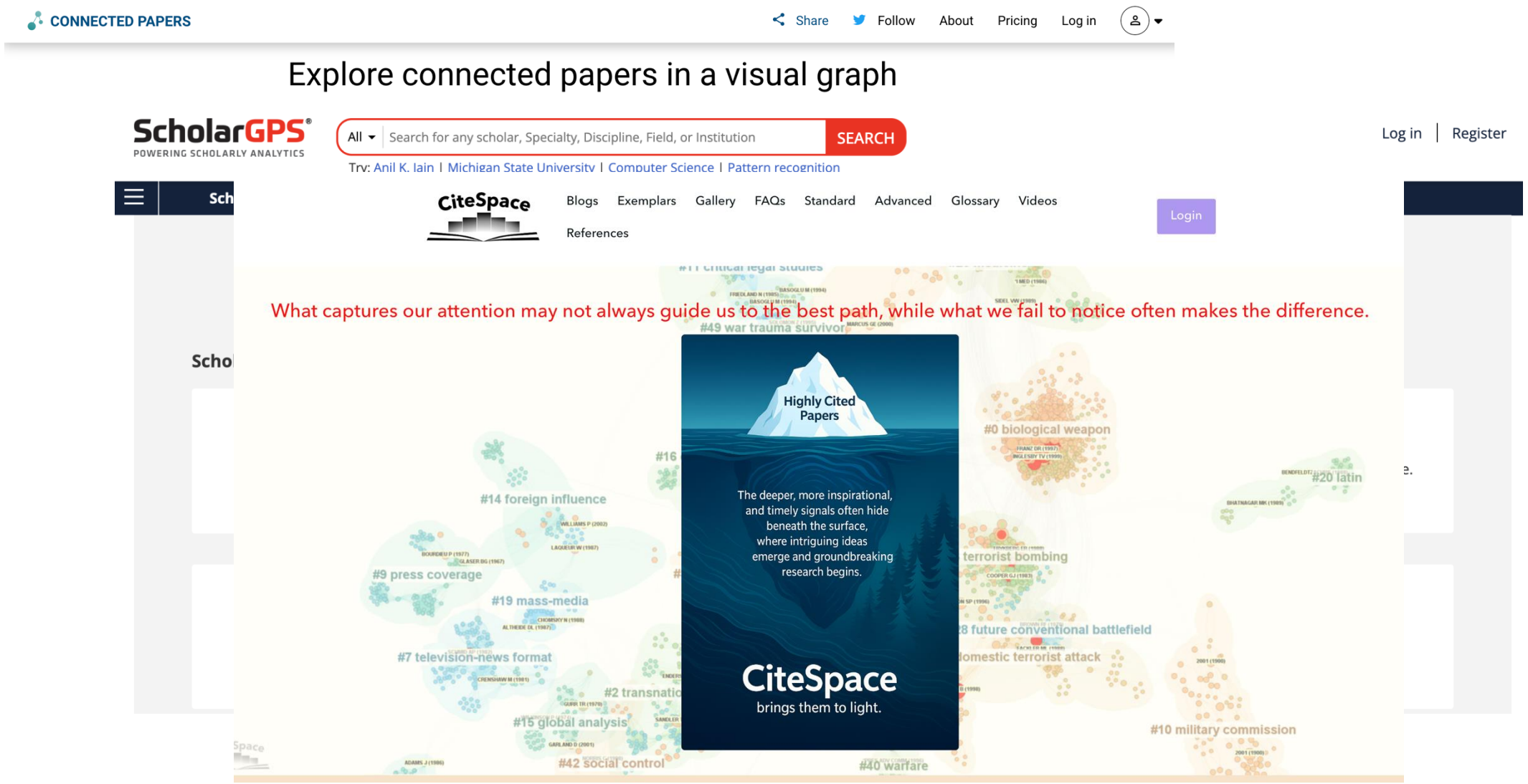
3 - Graph Based System



3 - Graph Based System

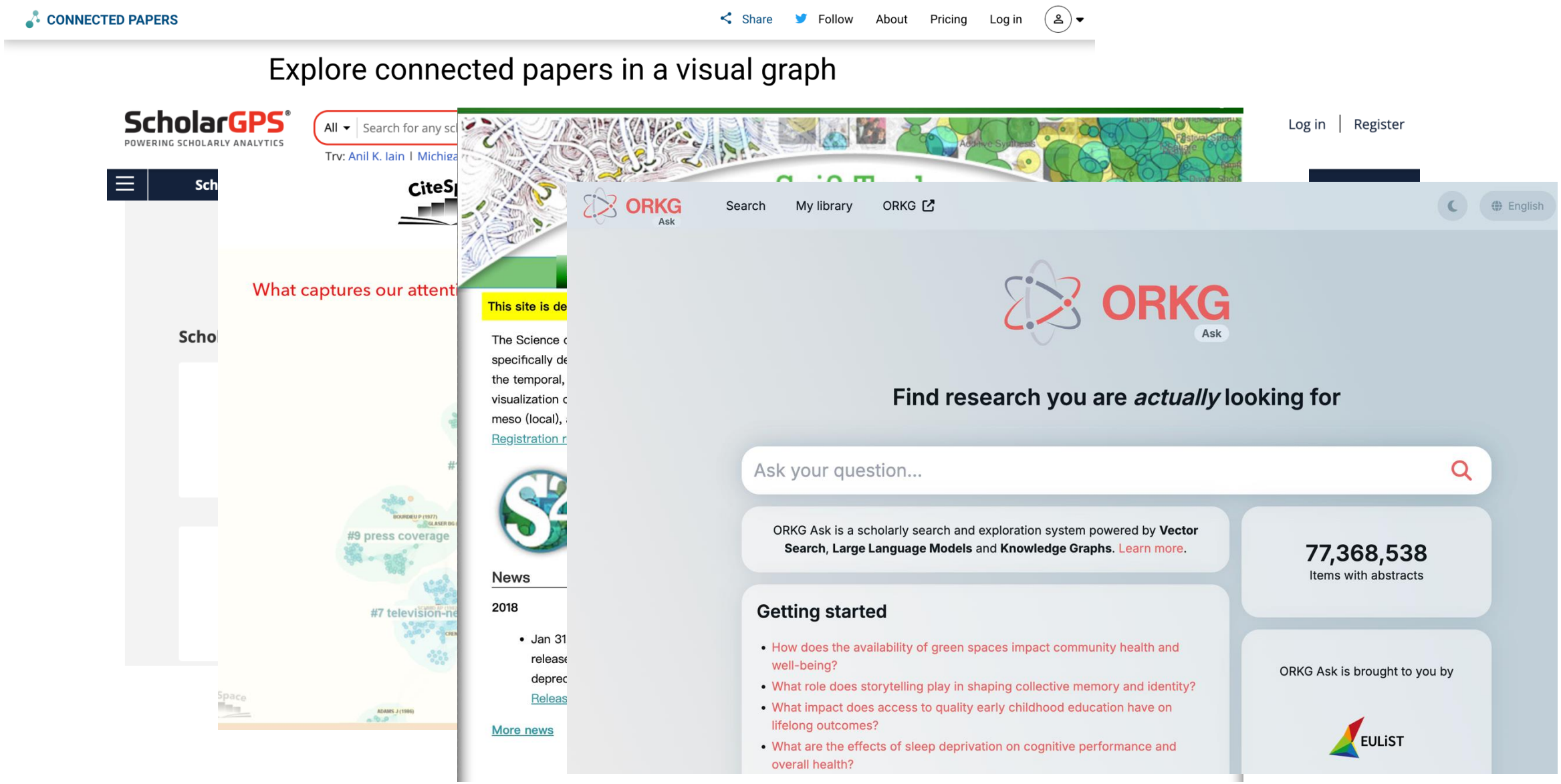


3 - Graph Based System

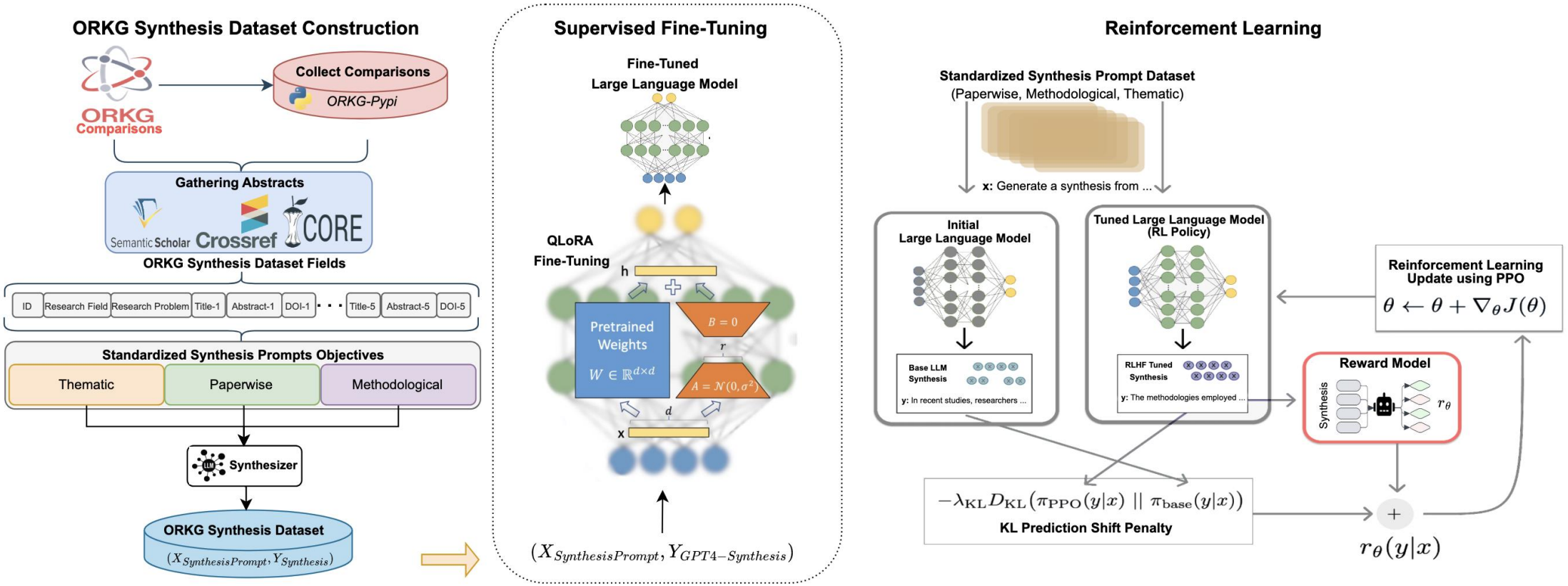


3 - Graph Based System

3 - Graph Based System



3 - Graph Based System



[LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis, 2024]

Performance and Conclusion

- SFT+RLAIF performs the best.

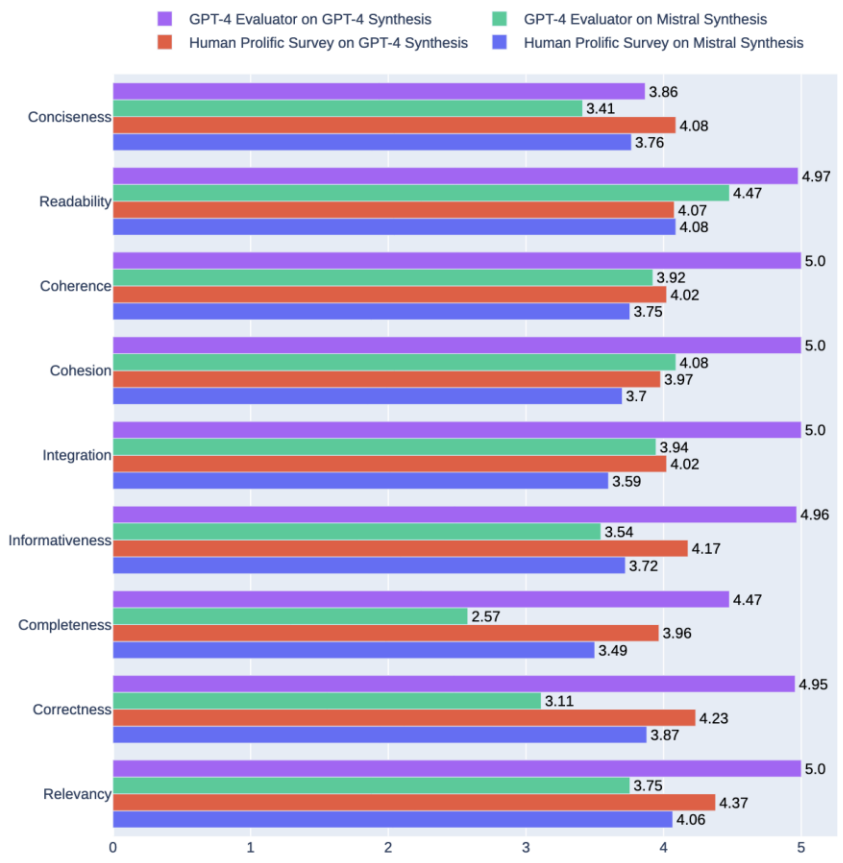
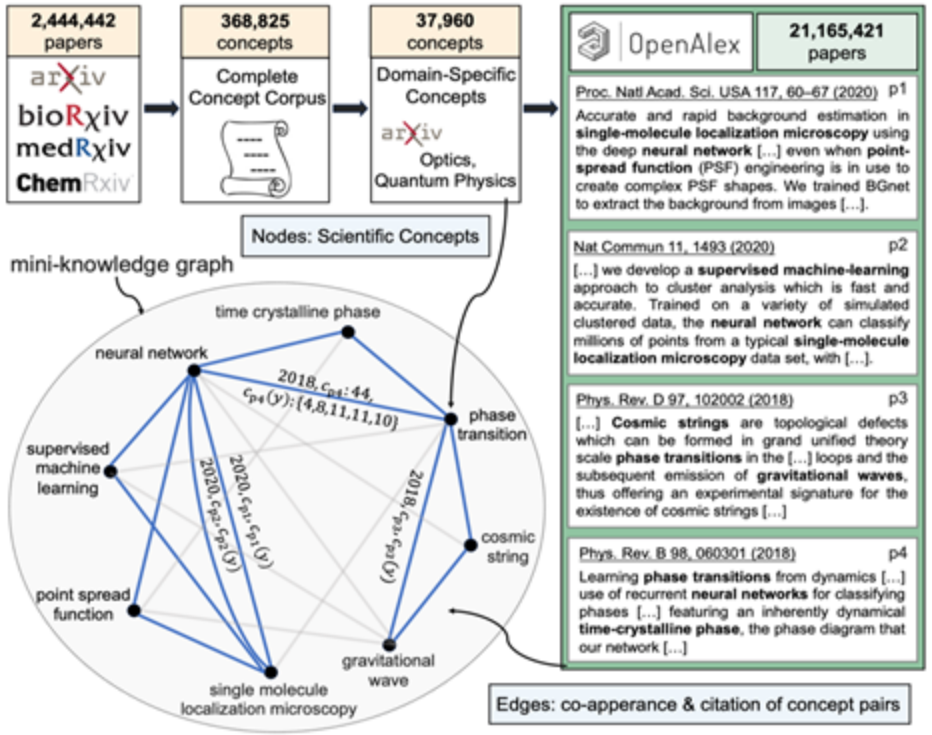
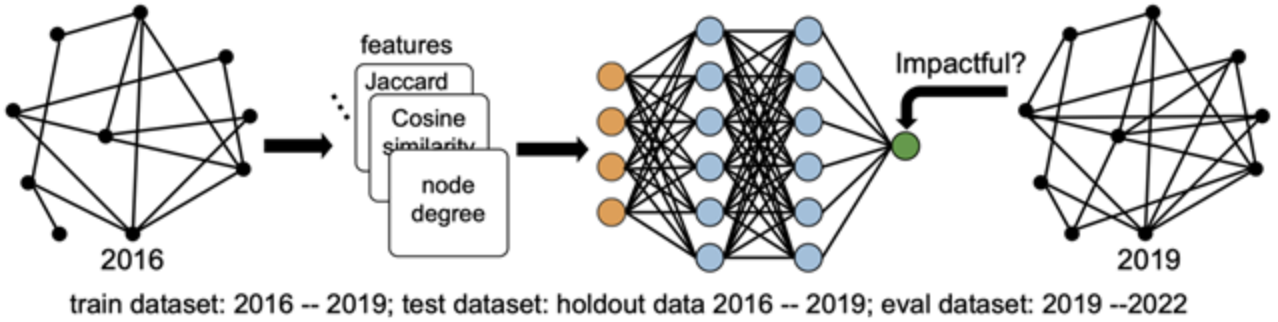


Figure 3: Consistency comparison of the GPT-4 evaluator between the *Vanilla* and *SFT+RLAIF (w/ GPT-4 Features)* models, assessed through three evaluations on the test set.

Trending Prediction

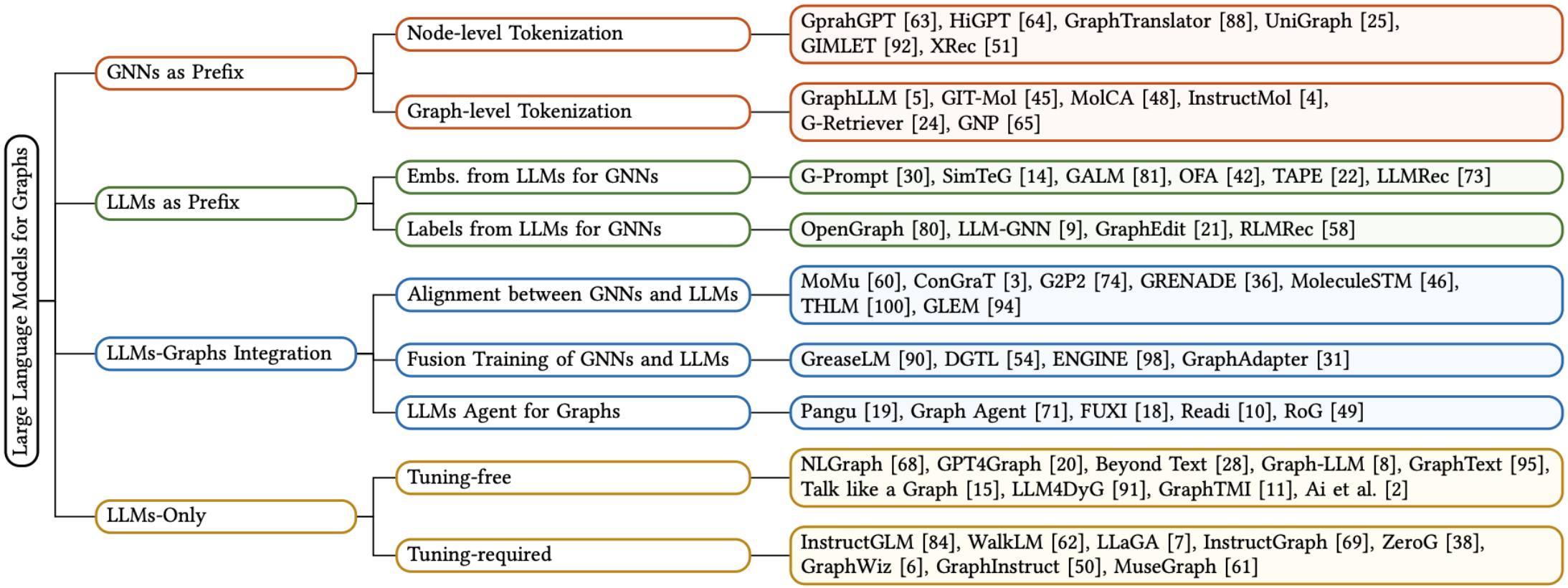
- Predict the impact of onsets of ideas.
- Extract 141 features for each pair are calculated.
  - 41 network features
  - 58 of these are node citation features
  - 42 features are about vertex pairs
- Graph Learning?





Forecasting high-impact research topics via machine learning on evolving knowledge graphs, 2025

LLMs + Graph?



A Survey of Large Language Models for Graphs, 2025

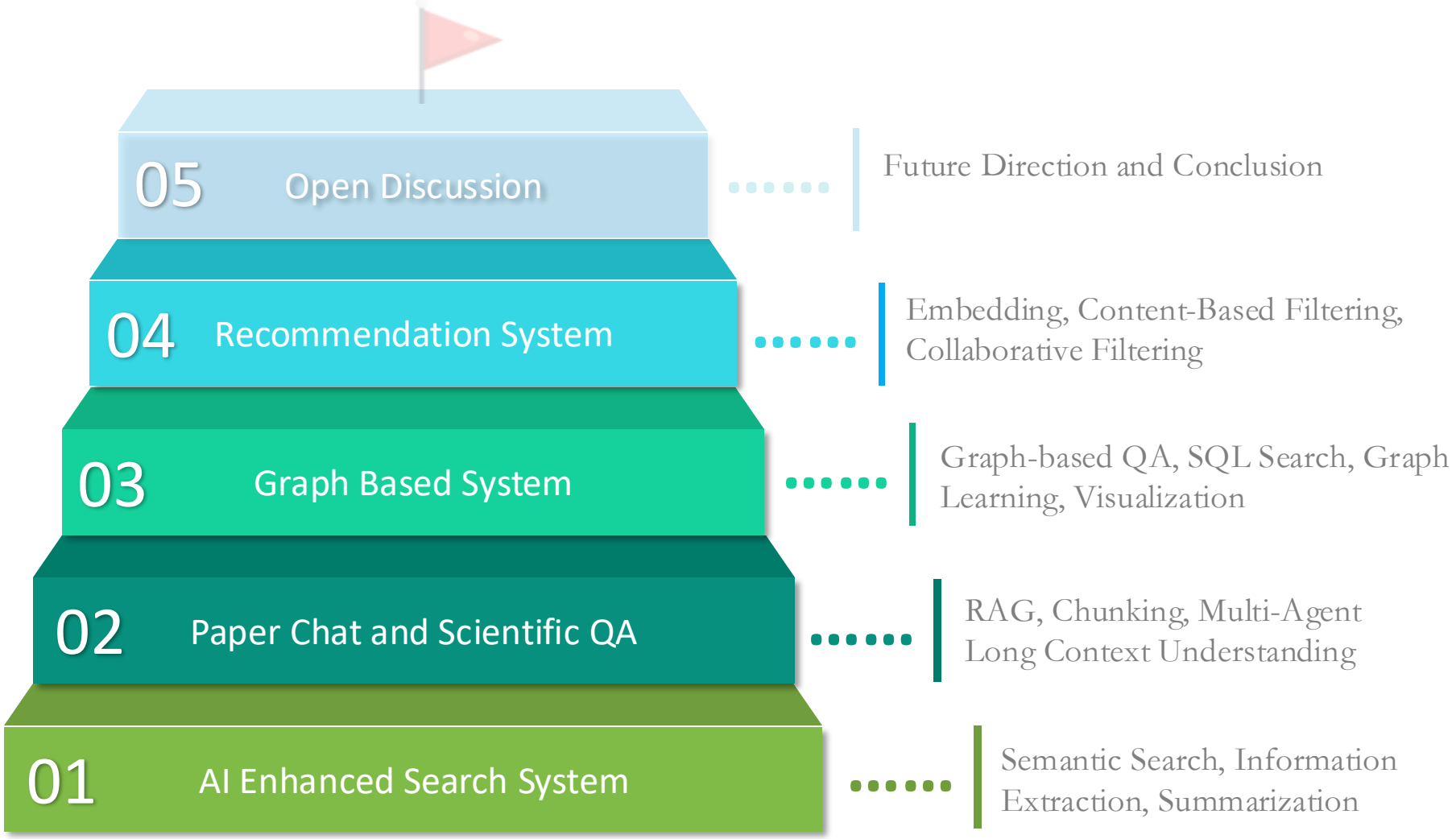
3 (Summary) - Graph Based System

- Main Function
  - Relationships between research papers
  - Explore knowledge structures

- Key techniques
  - Citation & Co-Authorship Networks
  - Graph Visualization & Navigation
  - Trending & Citation Analysis

- Challenges
  - Effective Graph Representation
  - Graph Update
  - Integration with LLMs

Outline

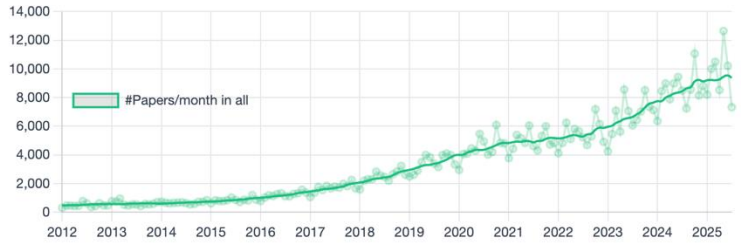

05
Open Discussion
Future Direction and Conclusion
04
Recommendation System
Embedding, Content-Based Filtering, Collaborative Filtering
03
Graph Based System
Graph-based QA, SQL Search, Graph Learning, Visualization
02
Paper Chat and Scientific QA
RAG, Chunking, Multi-Agent Long Context Understanding
01
AI Enhanced Search System
Semantic Search, Information Extraction, Summarization

4 - Recommendation System

Papers Communities Organizations Events Blog Pricing Search Sign up Login

Stay Updated on the Trends, Connect with AI Researchers.

#arXiv papers on AI



Upcoming ResearchTrend Connect Sessions

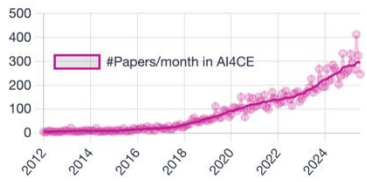
Date	Community	Location	Event

No events available

Feedback

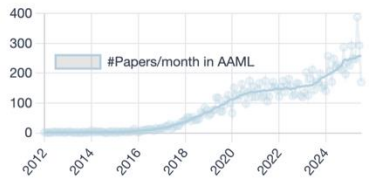
All communities

AI4CE AI for Computational... Follow 3



Papers 14.6K # Contributors 46.6K # followers 3

AAML Adversarial Attacks in... Follow 42

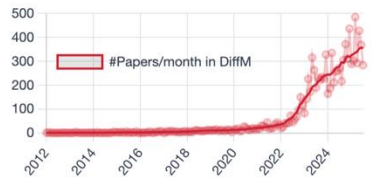


Papers 13.9K # Contributors 31.4K # followers 42

VLM Vision-Language Models Follow 150

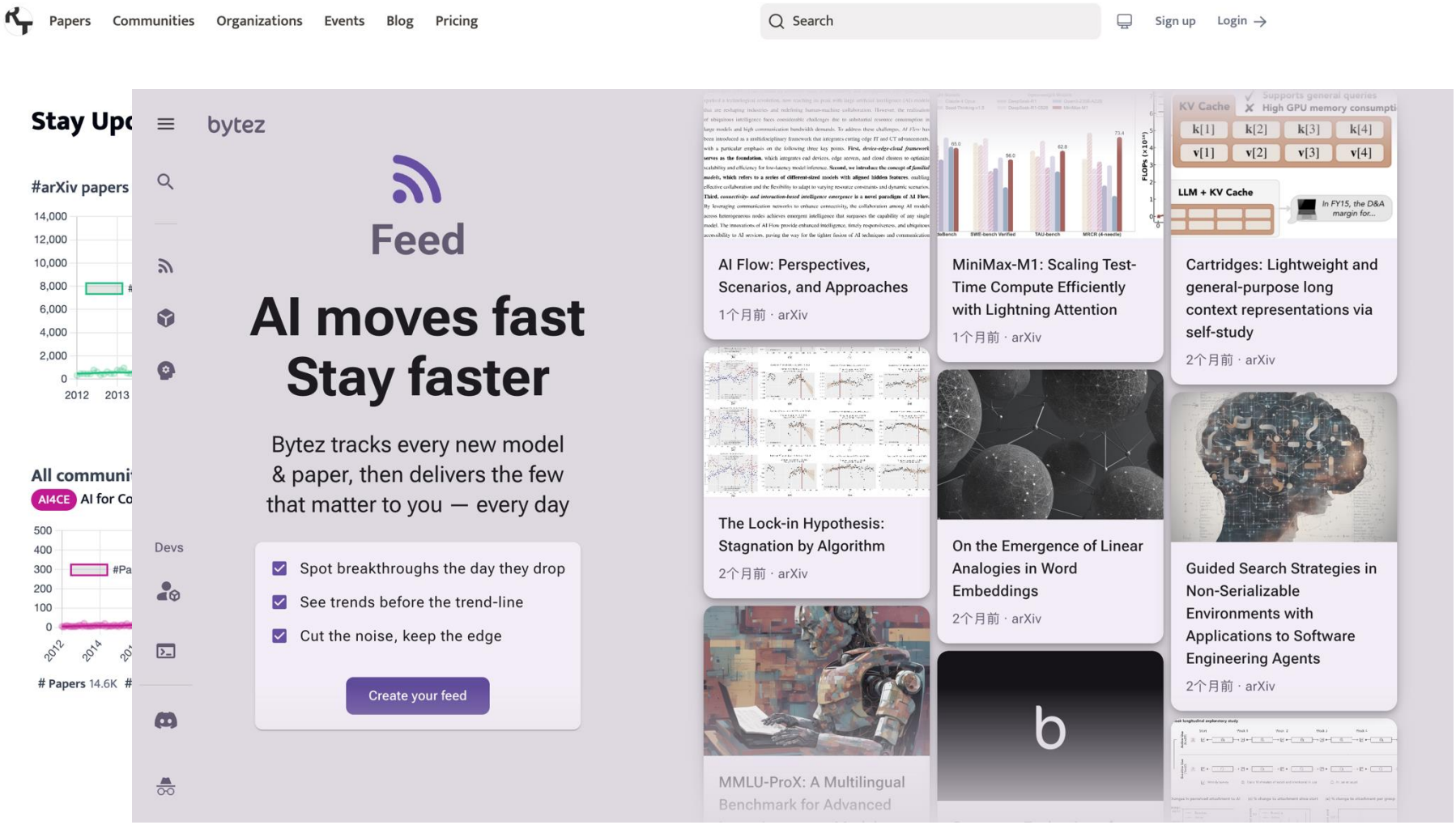
Papers 12.0K # Contributors 38.9K # followers 150

DiffM Diffusion Models Follow 116

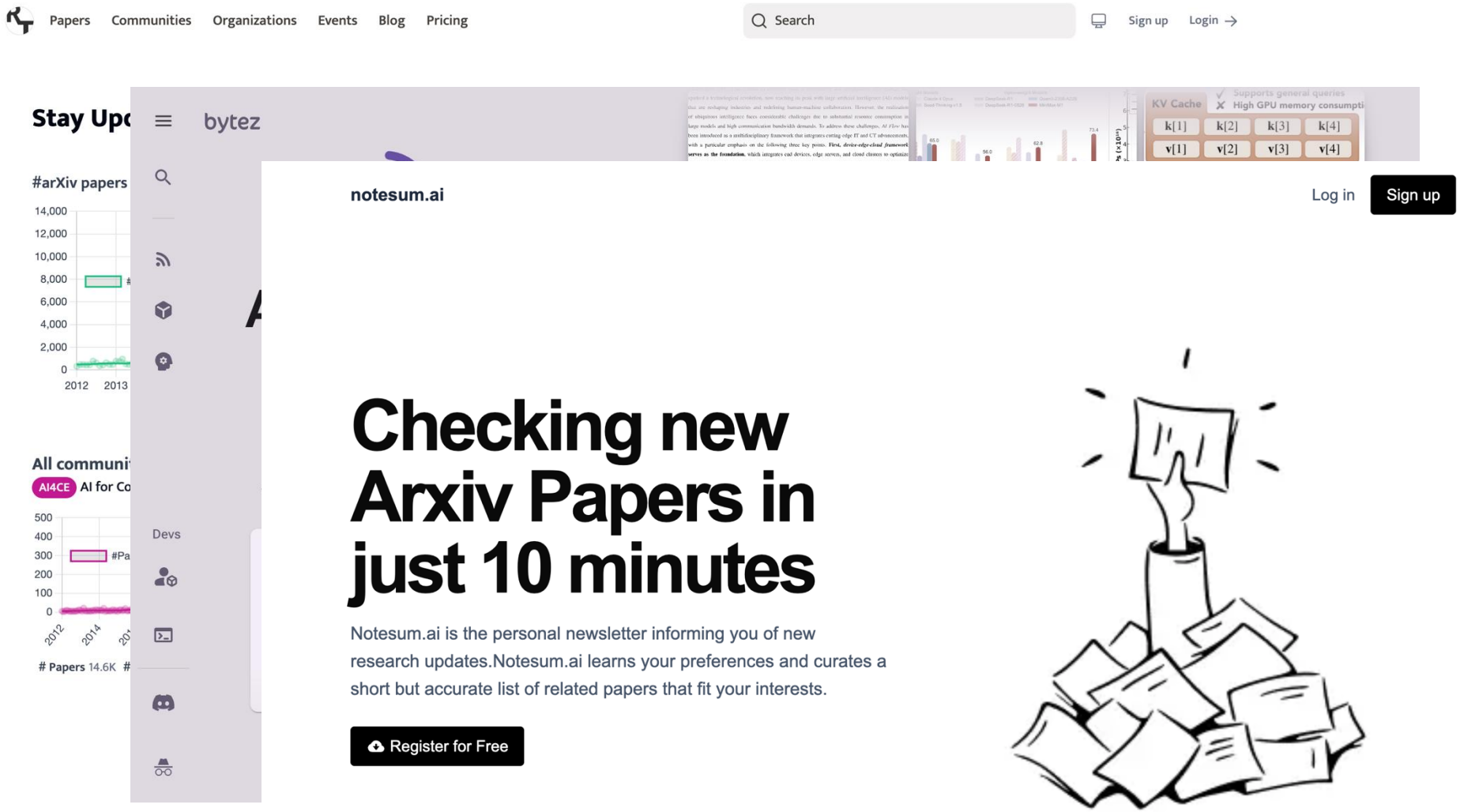


Papers 10.1K # Contributors 27.9K # followers 116

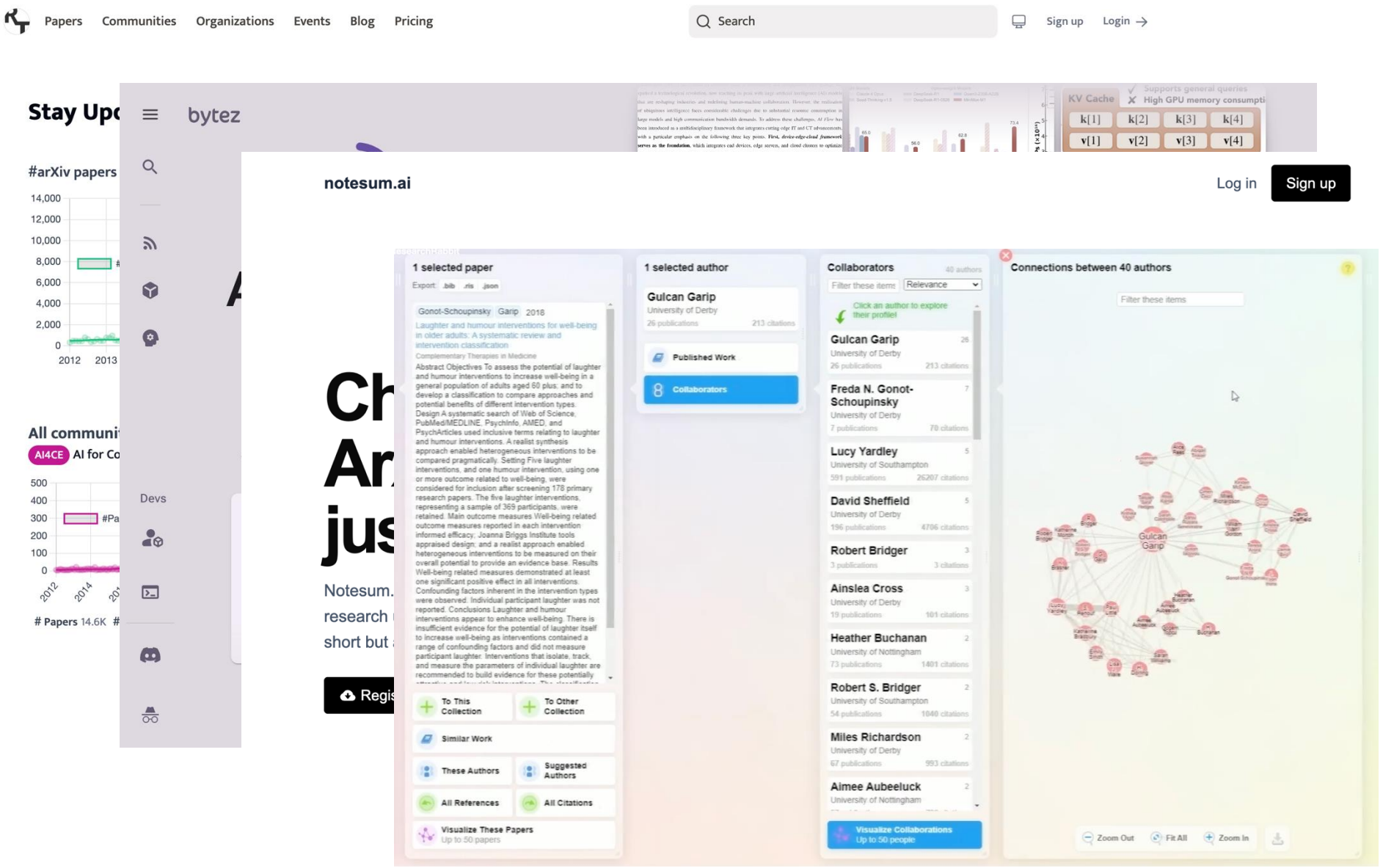
4 - Recommendation System



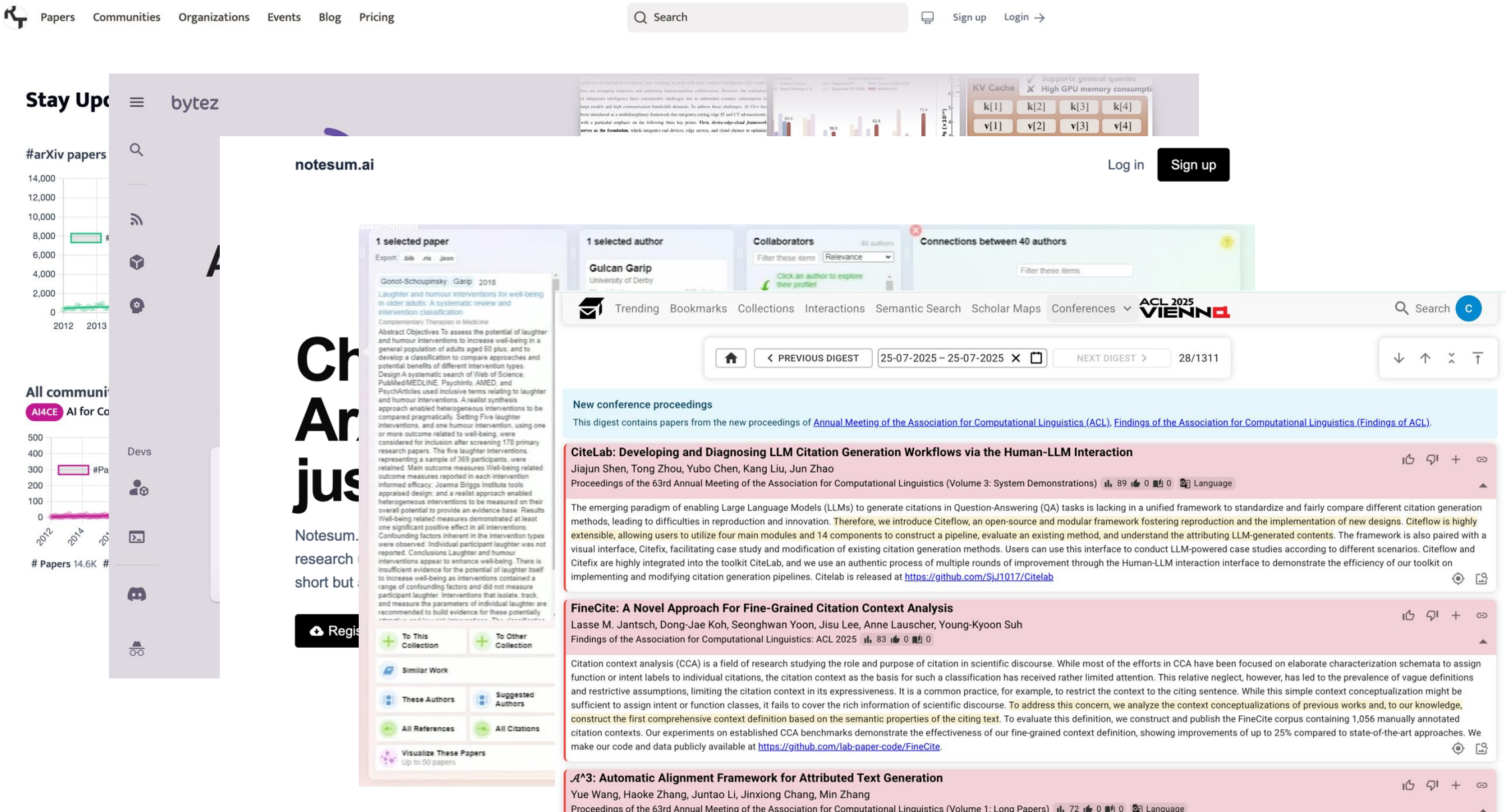
4 - Recommendation System



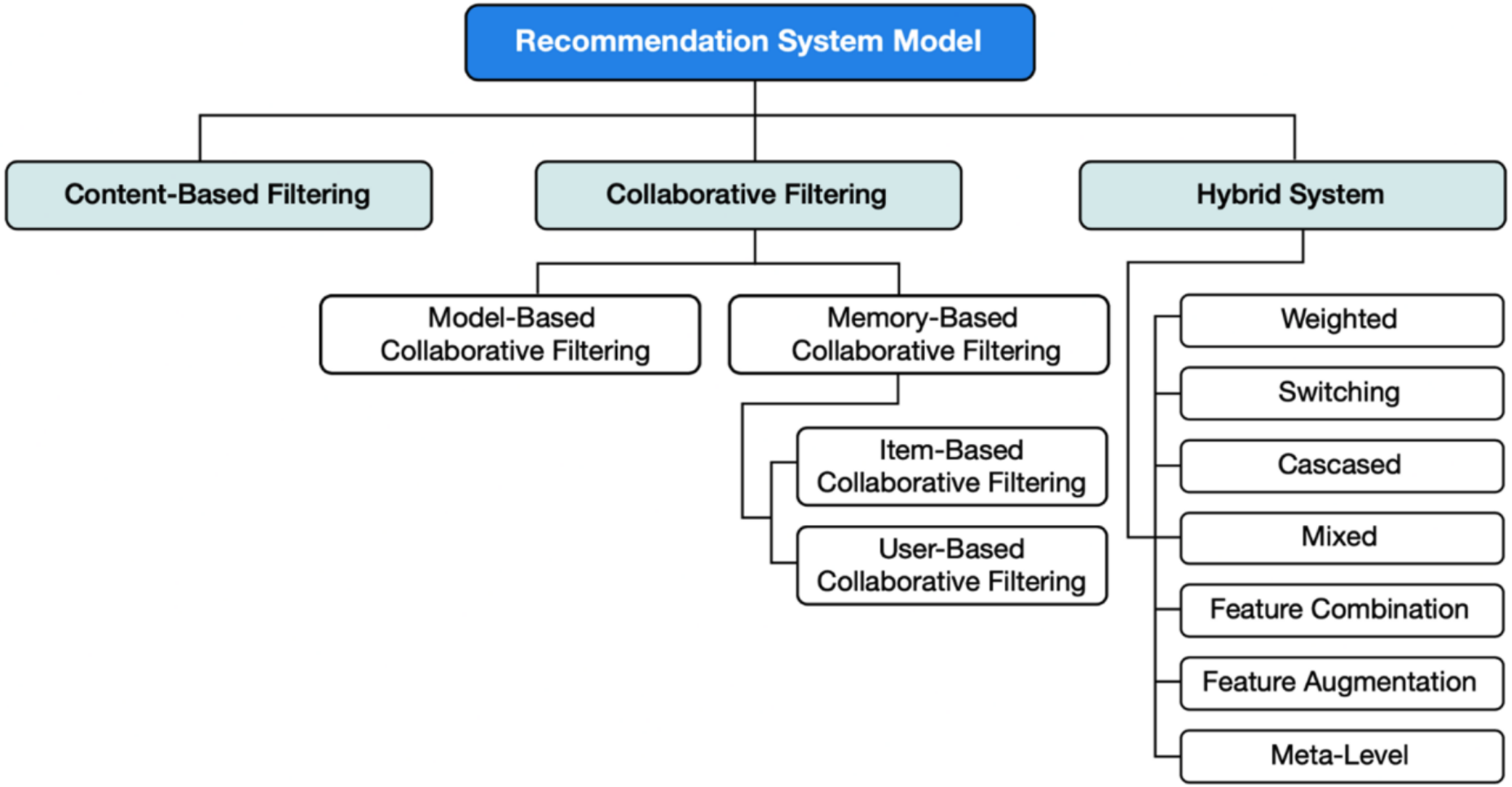
4 - Recommendation System



4 - Recommendation System

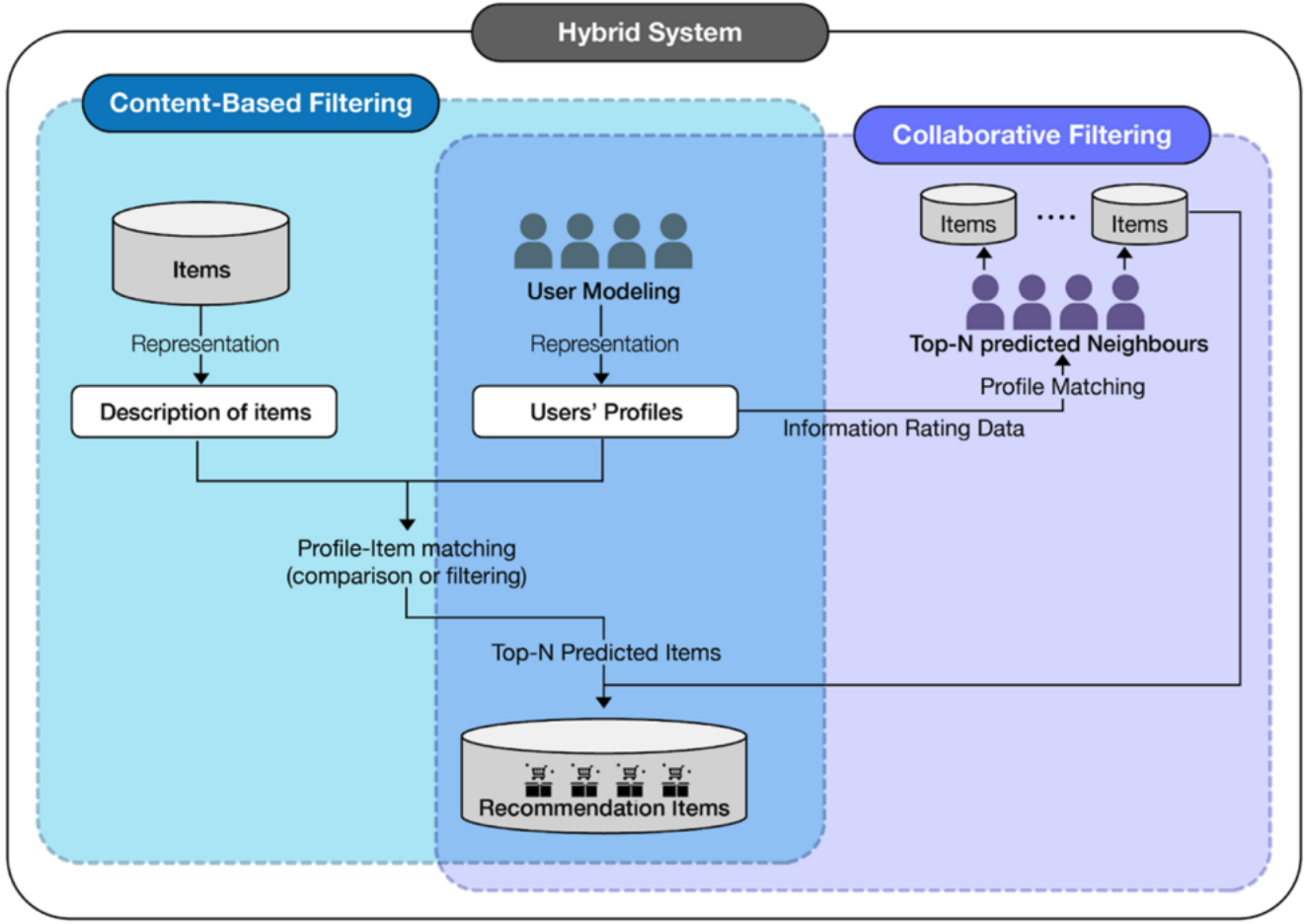


Recommendation System



A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields, 2022

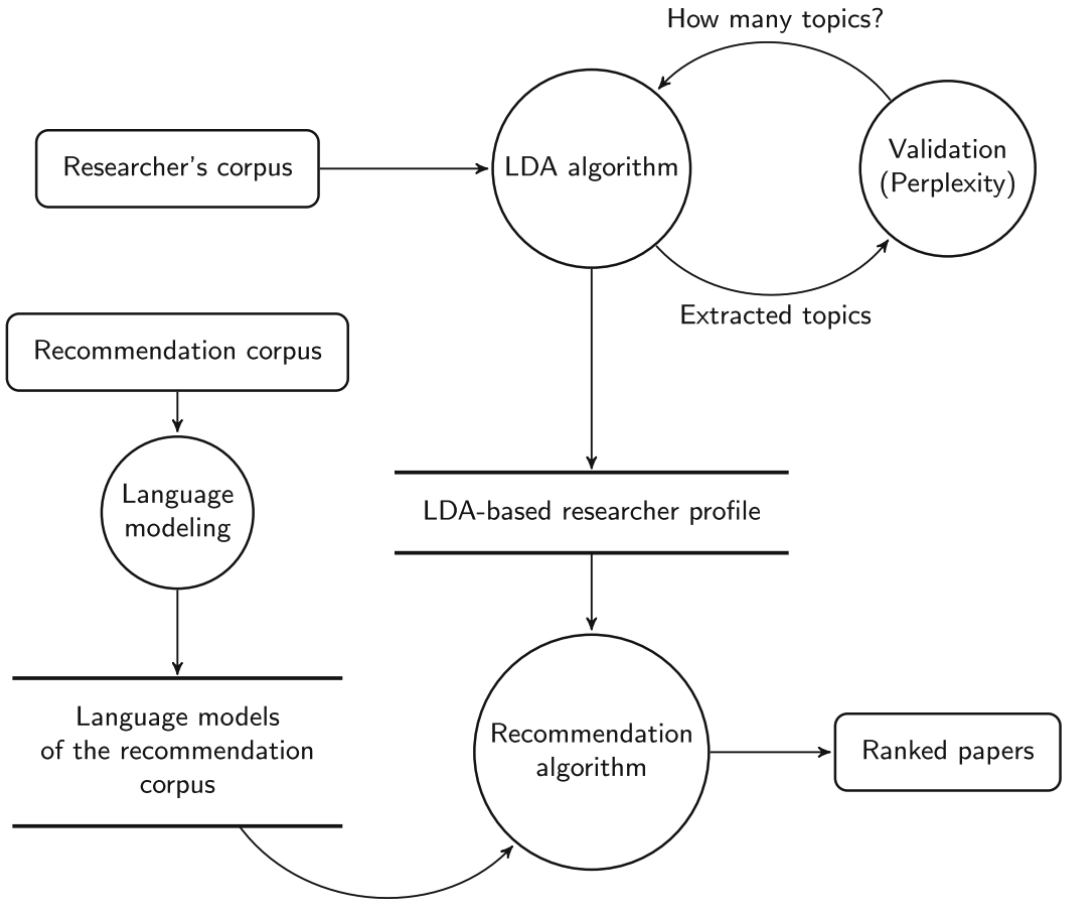
Pipeline of recommendation systems



A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields, 2022

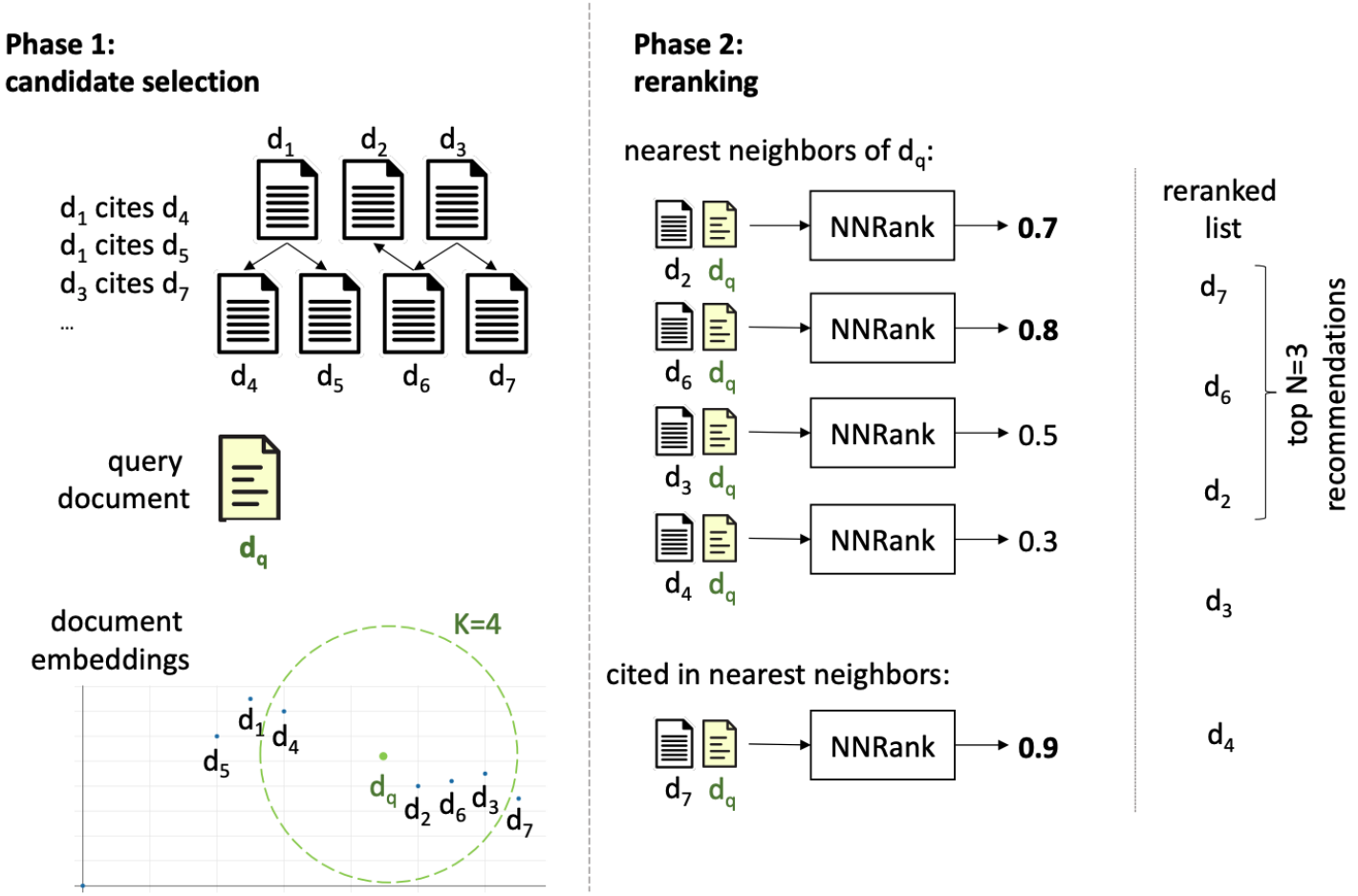
Content-based Methods

- Compares topics from the researcher's profile with the language models of unseen papers.
- Uses the symmetrized Kullback-Leibler divergence to measure similarity between probability distributions (topics and language models).
- **Limitations**
  - Shifted Cold-Start Problem
  - Potential for Limited/General Concepts
  - Lack of Contextual Citation Information



An LDA-Based Approach to Scientific Paper Recommendation, 2016

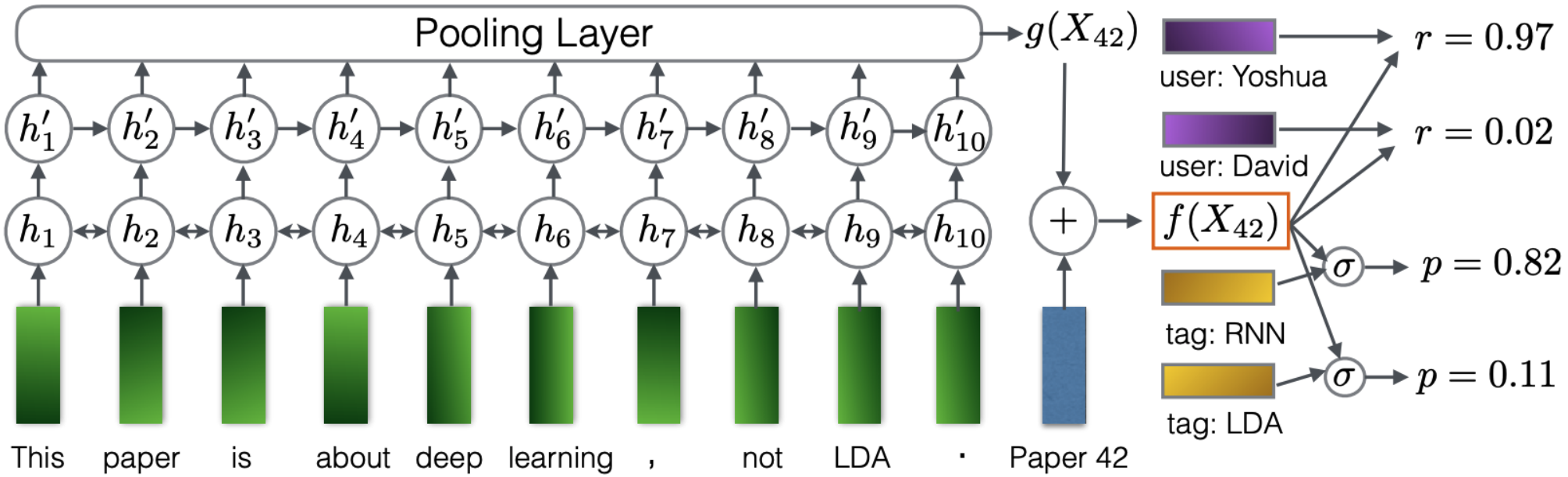
NNSelect+NNRank



[Content-Based Citation Recommendation, 2018]

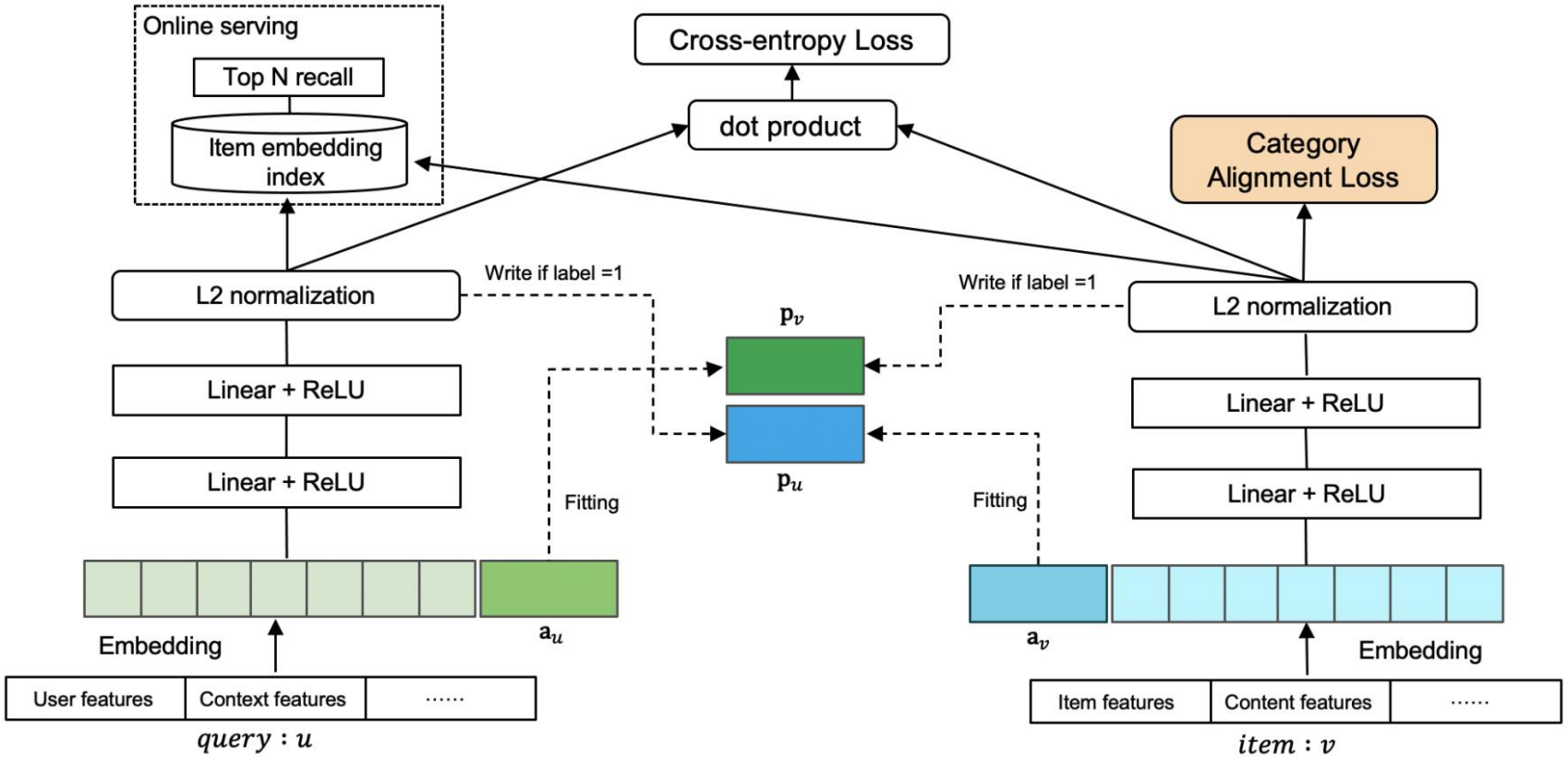
Collaborative Filtering

- By training a text encoder network as a combination of content recommendation and item metadata prediction (e.g., tag prediction)



Ask the GRU: Multi-Task Learning for Deep Text Recommendations, 2016

Hybrid Systems

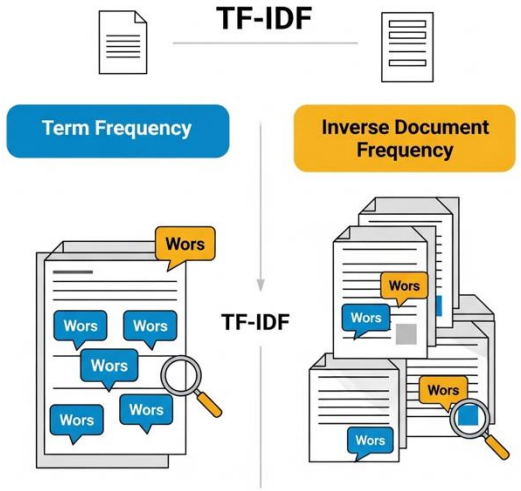


A Dual Augmented Two-tower Model for Online Large-scale Recommendation, 2021

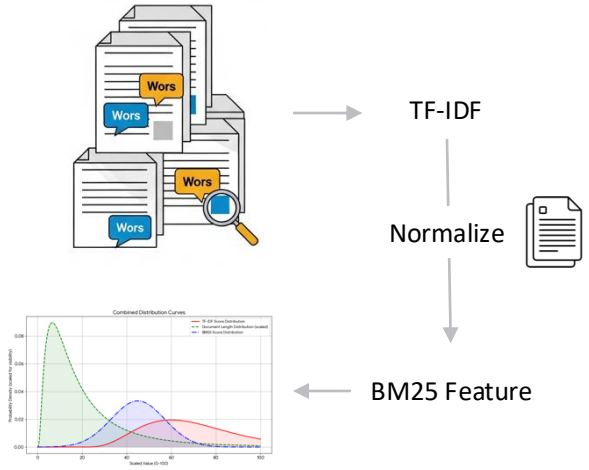
Comparsion of Recommendation Systems

Model Type	Pros	Cons
Content-Based Filtering	• Works well for new users • Personalized to individual preferences • Doesn’t require large user base • Easy to interpret recommendations	• Needs detailed item features • Limited discovery of new interests
Collaborative Filtering	• Doesn’t need item metadata • Leverages collective user behavior	• Struggles with sparse data • Scalability issues in memory-based
Hybrid Systems	• Combines best of all models • Mitigates individual weaknesses • Often improves accuracy	• Increased complexity • Requires more computation

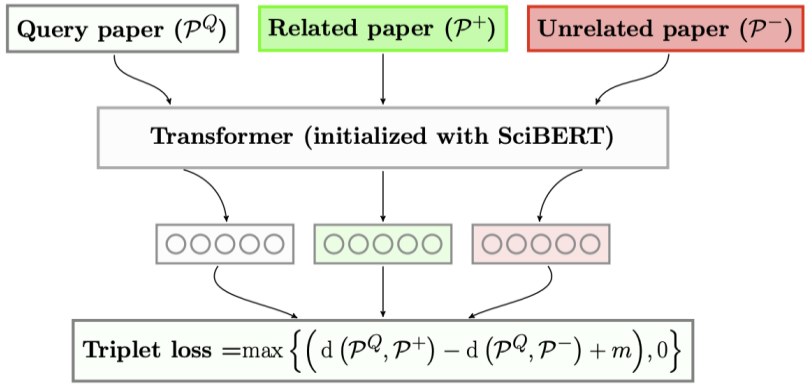
Document Embeddings



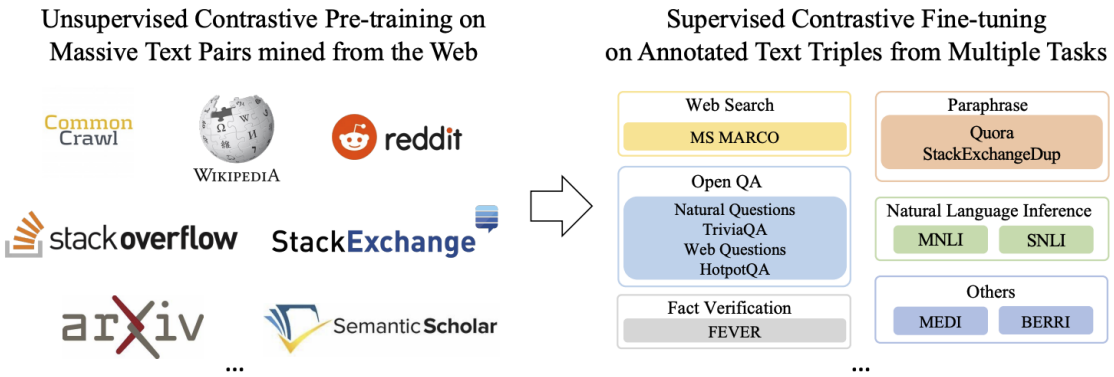
TF-IDF



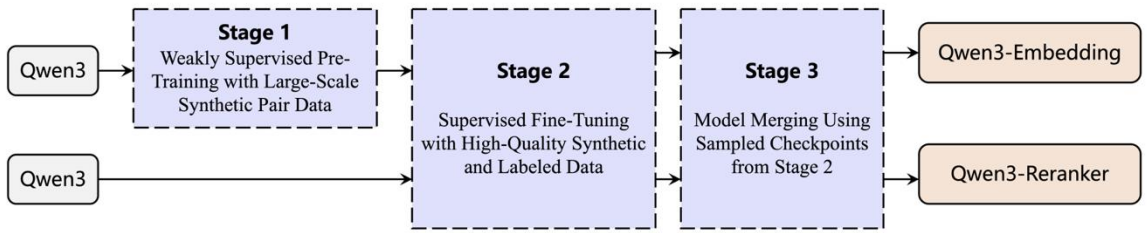
BM25



SPECTER



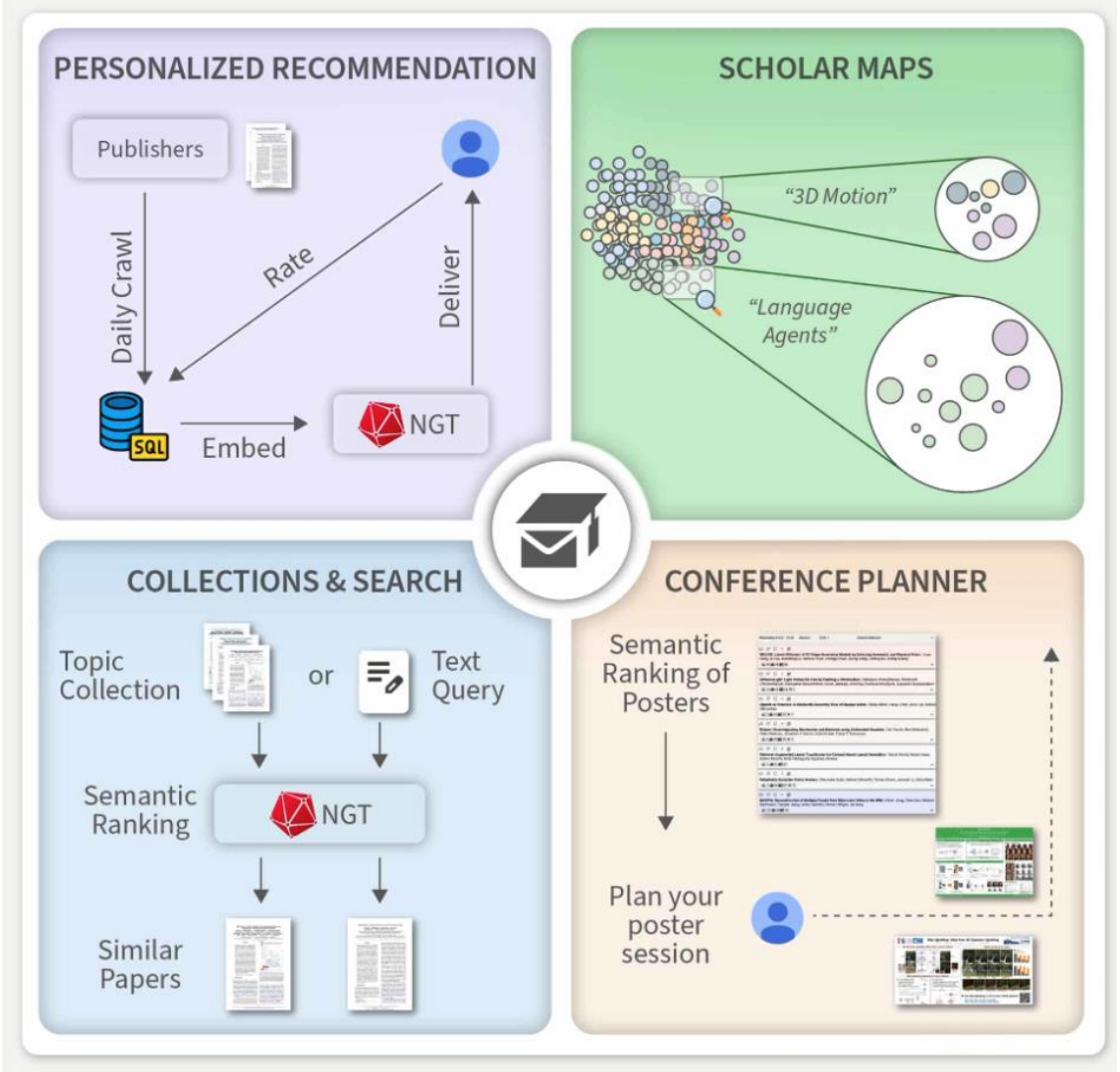
GTE



Qwen-3 Embedding

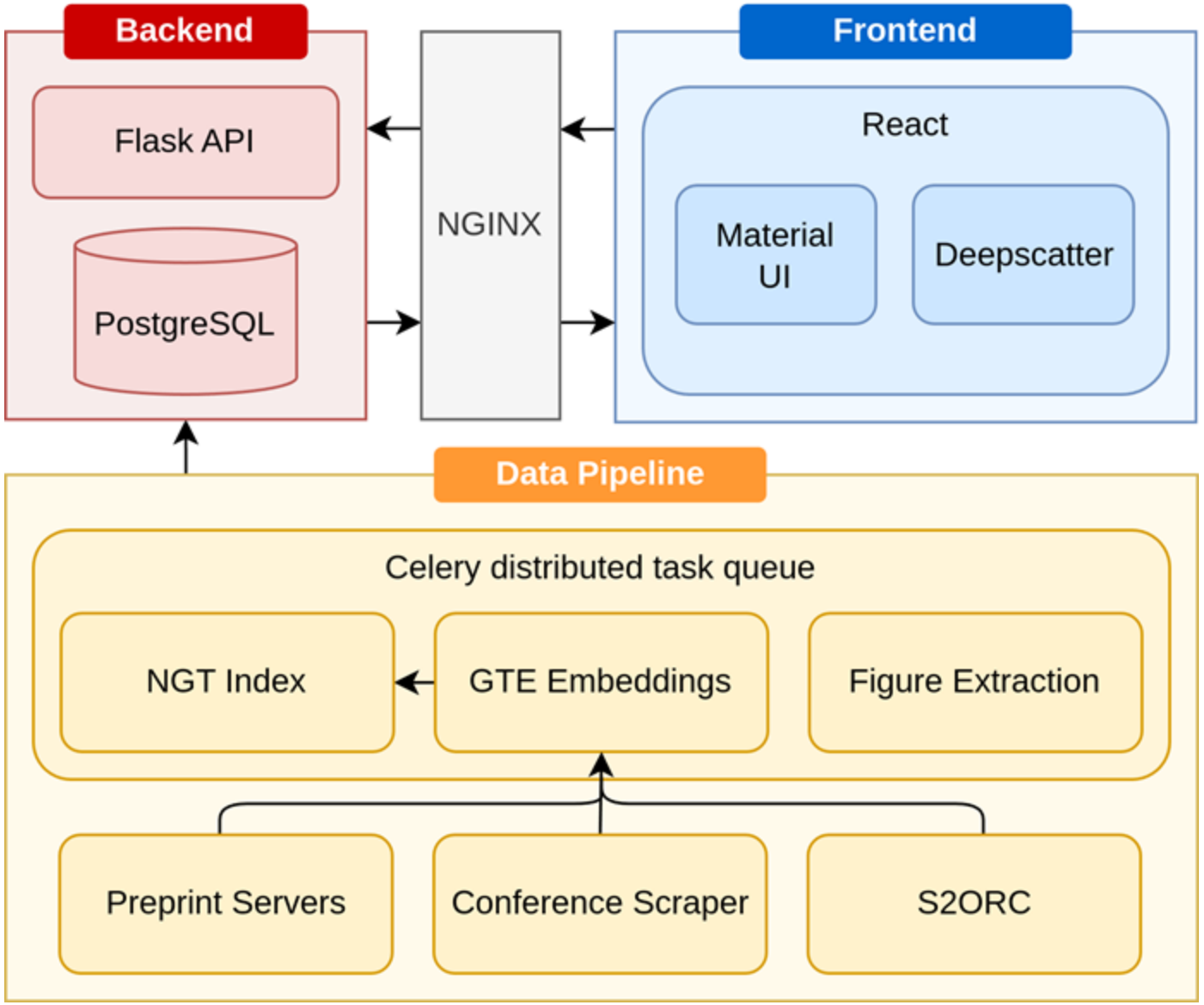
Scholar Inbox – 2025 Demo

- Personalized Recommendations
- Scholar Maps
- Collections
- Conference Planner



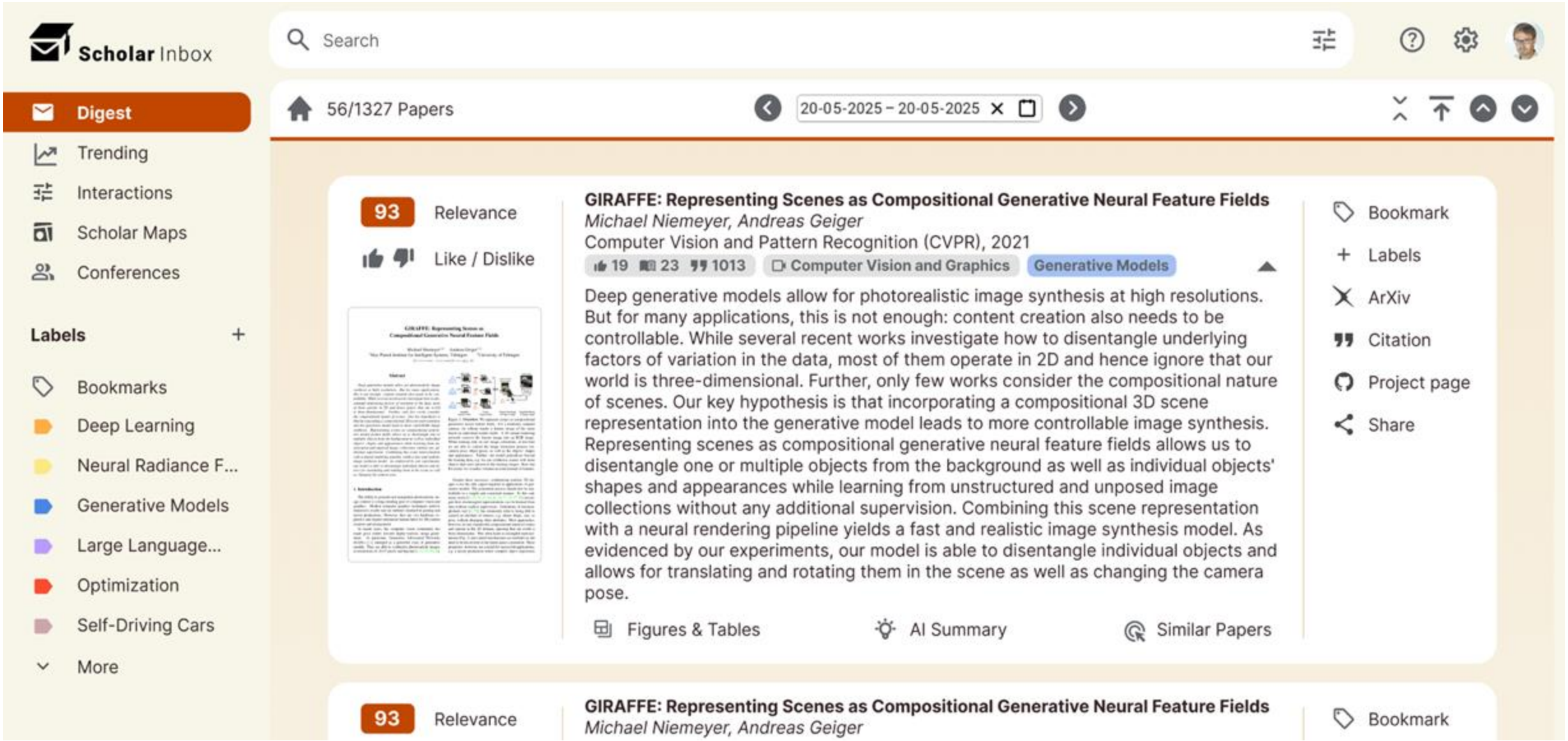
[Scholar Inbox: Personalized Paper Recommendations for Scientists, 2025]

Software Flow

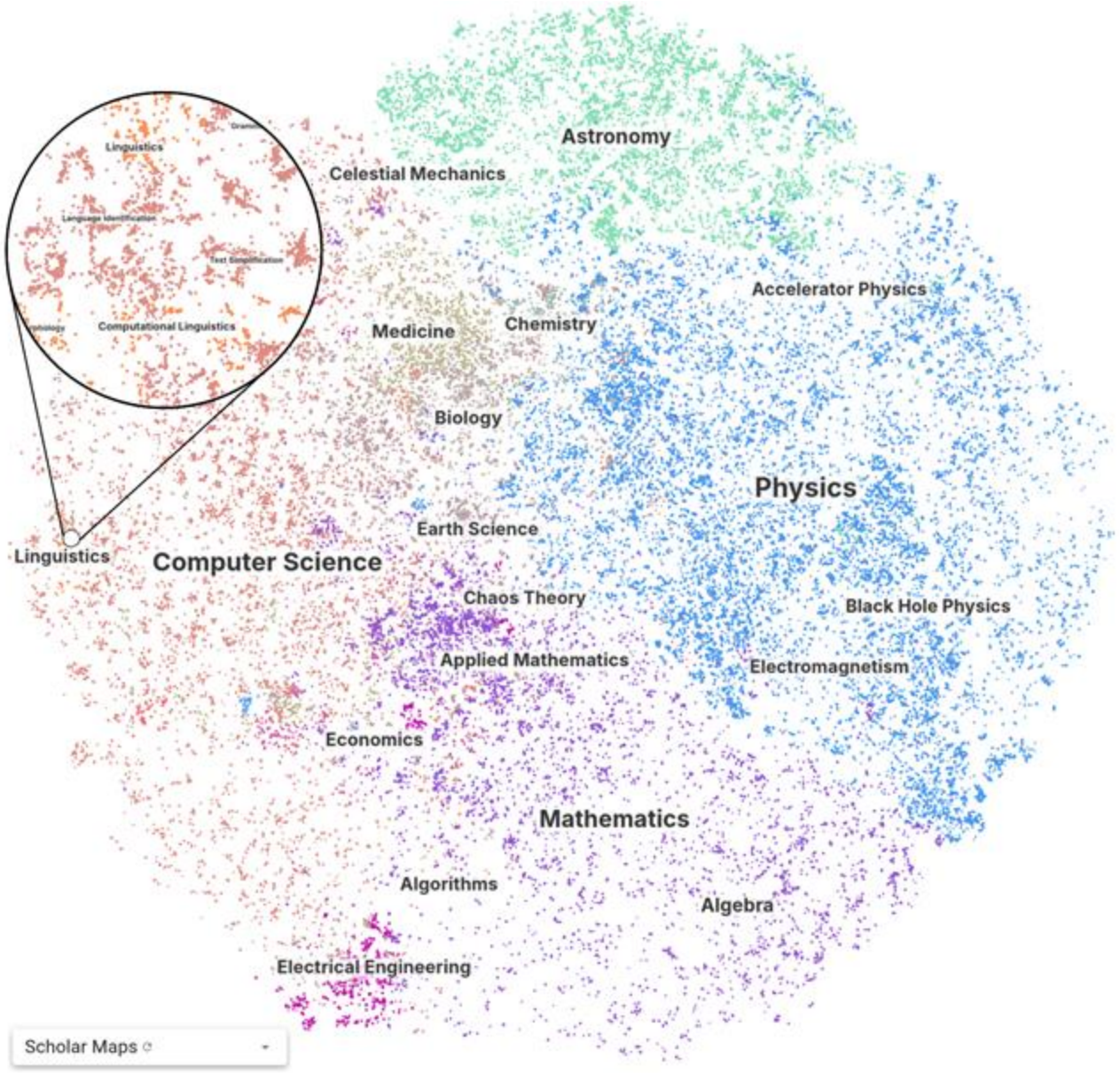


Scholar Inbox: Personalized Paper Recommendations for Scientists, 2025

Recommender system

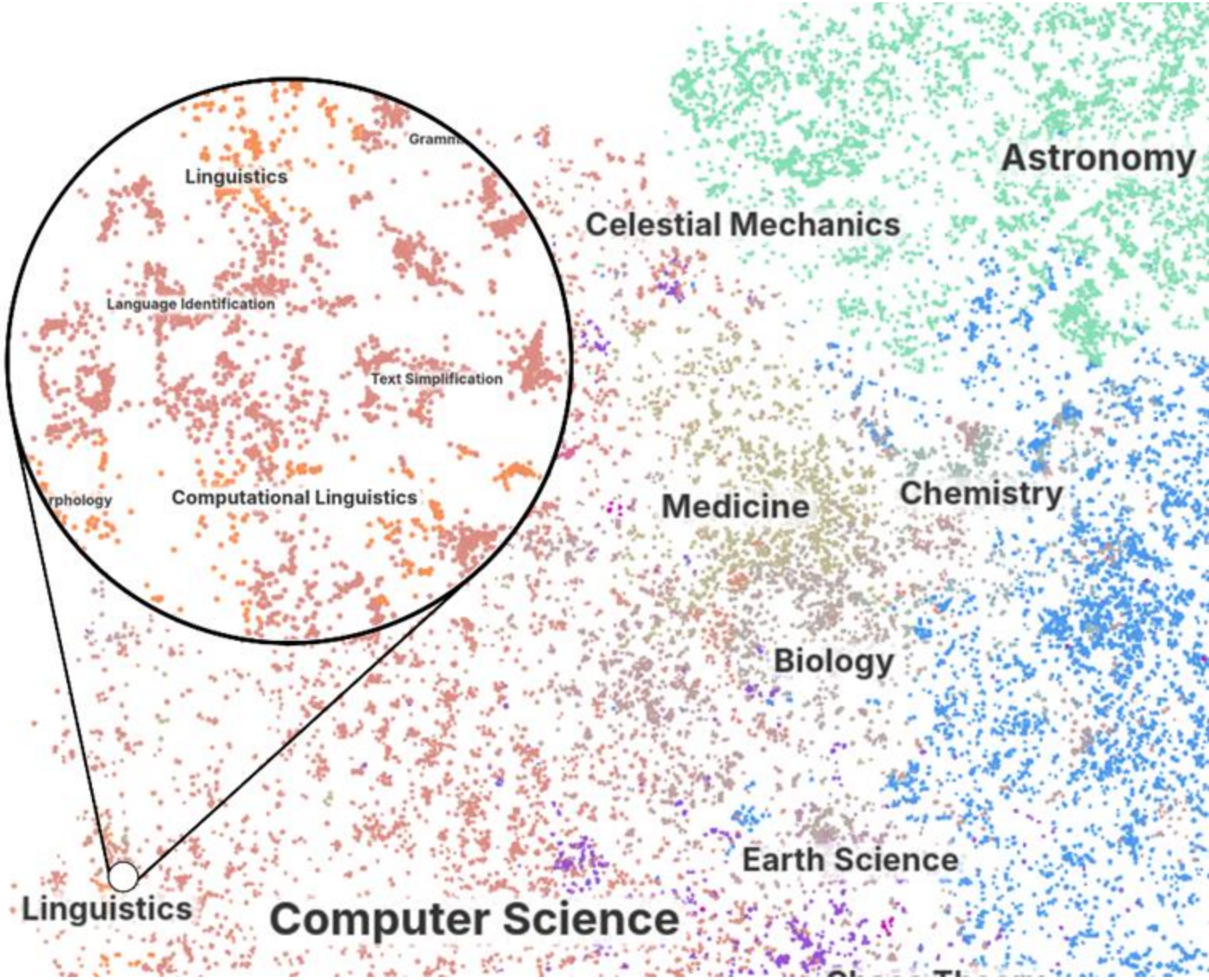


Scholar Maps



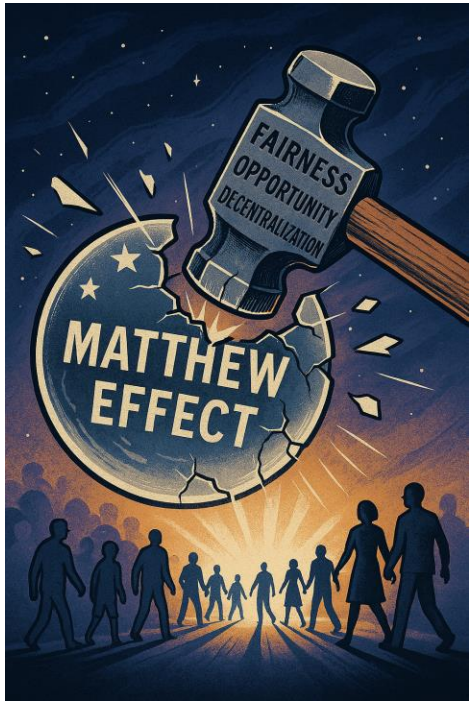
Scholar Inbox: Personalized Paper Recommendations for Scientists, 2025

Scholar Maps

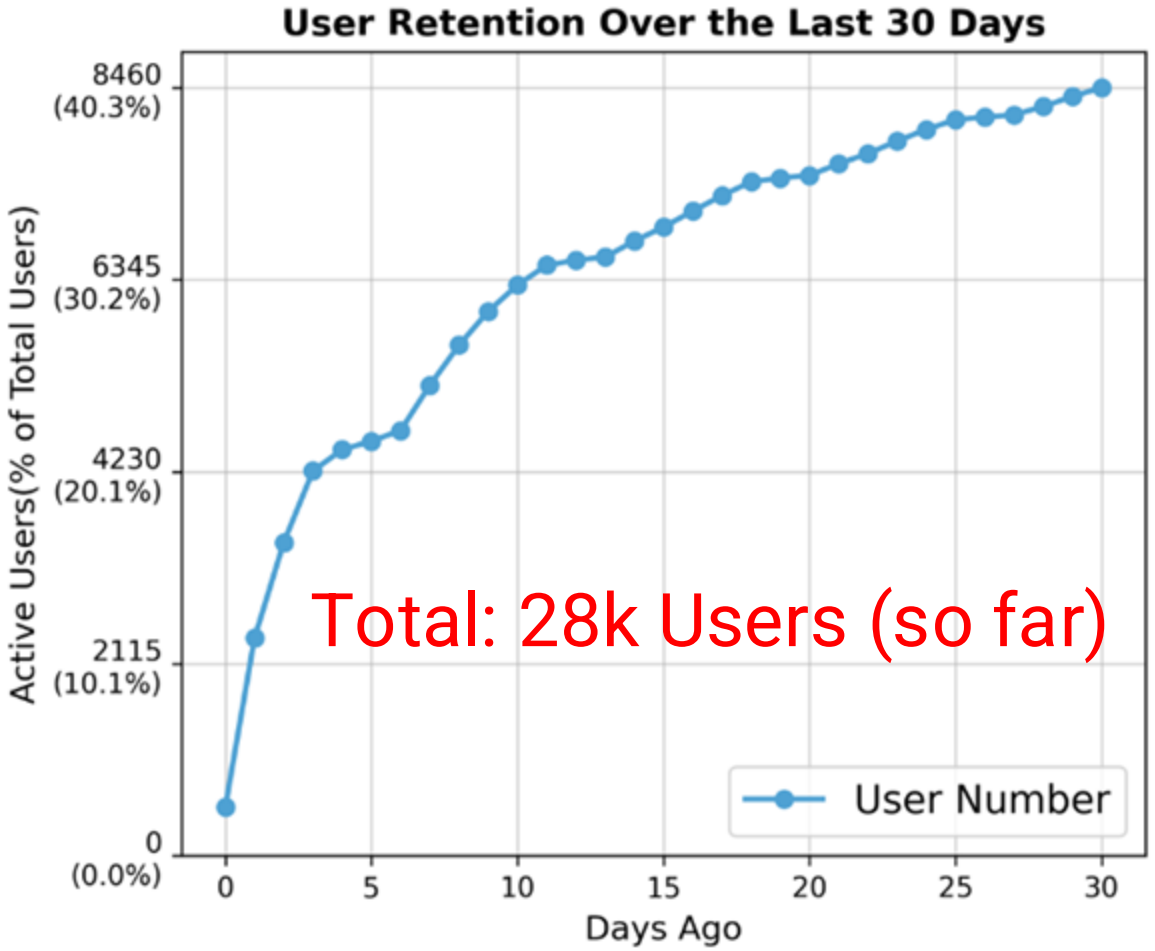


Scholar Inbox: Personalized Paper Recommendations for Scientists, 2025

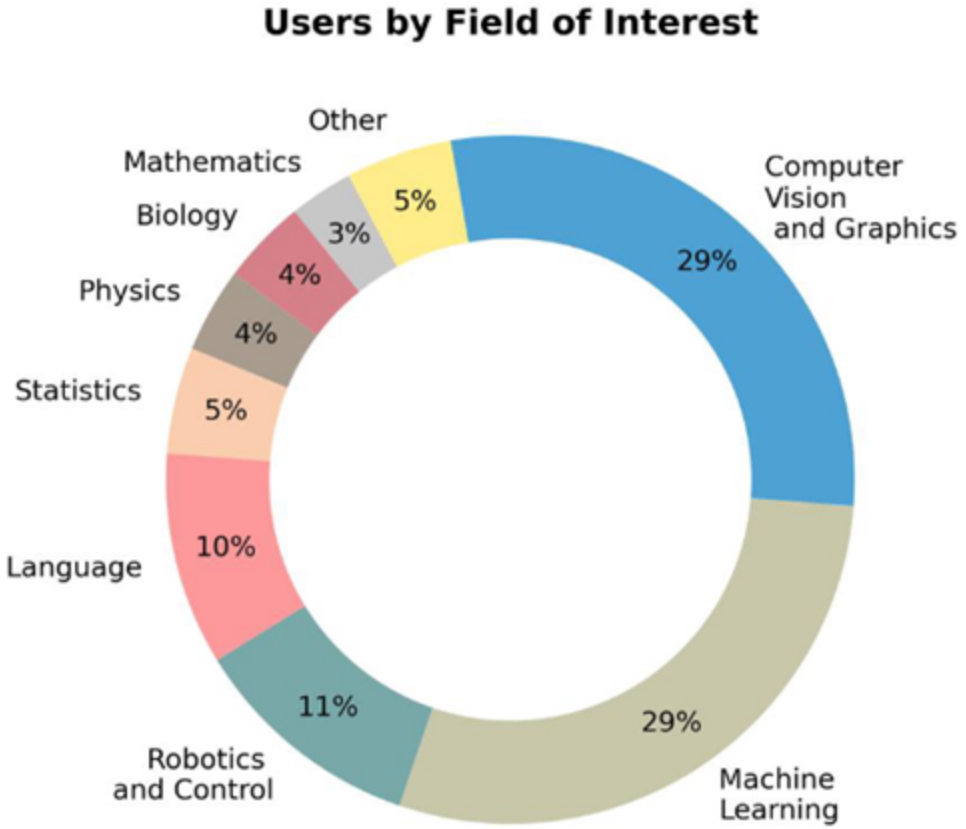
Why we create Scholar Inbox?



9.000 users come back again and again



(a) User Retention



(b) User Domain Distribution

Welcome to check our paper

- First time serving NLP conference.
- Hall 5X, 28th July, 11:00-12:30, ACL 2025.

Scholar Inbox supports ACL25

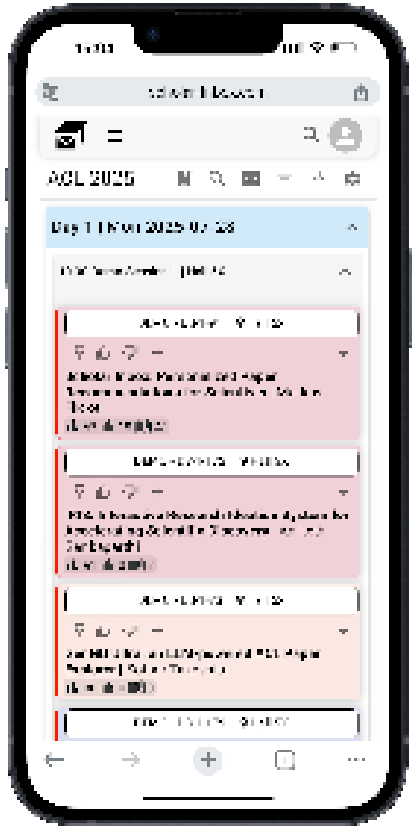
With your personalised relevance:

- Explore the ACL25 timetable
- Prioritise within poster sessions
- Plan your visit

Make this year your most efficient ACL visit ever!

Join our 30k users

www.scholar-inbox.com



Scholar Inbox: Personalized Paper Recommendations for Scientists

Markus Flicke Glenn Angrabeit Madhav Iyengar Vitalii Protsenko Illia Shakun Jovan Cicvaric Bora Kargi Haoyu He Lukas Schuler Lewin Scholz Kavyanjali Agnihotri Yong Cao Andreas Geiger

University of Tübingen, Tübingen AI Center

www.scholar-inbox.com


NEVER
MISS
A BEAT
AGAIN.

4 (Summary) - Recommendation System

- Main Function
  - Keep updated literatures
  - Personlized Recommendation
  - Collections

- Key techniques
  - Content-Based Filtering
  - Collaborative Filtering
  - Semantic Search & Embeddings

- Challenges
  - Cold Start Problem
  - Overpersonalization vs Matthrew Effect
  - Dynamic Interests of Researchers

Overview of popular literature search

	Platform	Search	Recommendations	Collections	Citation Analysis	Trending Analysis	Author Profiles	Visualization Tools	Paper Chat	Idea Generation	Paper Writing	Summarization	Paper Review	Datasets	Code Repositories	LLM Integration	Web API	Personalization	Cost	Data Source
AI-Enhanced Search	Elicit	✓						✓	✓		✓	✓			✓				Freemium	125 million
	OpenScholar	✓		✓				✓			✓				✓				Free	45 million
	Undermind	✓		✓				✓			✓				✓			✓	Premium	over 200 million
	Perplexity	✓						✓	✓		✓	✓			✓				Freemium	
	Consensus	✓		✓				✓			✓				✓	✓			Freemium	over 200 million
	SciSpace	✓		✓				✓	✓		✓	✓			✓				Freemium	
	scienceQA	✓		✓	✓			✓	✓		✓	✓			✓				Freemium	220 million
	PaperQA2							✓						✓	✓				Free	
	Paperguide	✓		✓				✓	✓		✓	✓			✓				Freemium	
	HyperWrite	✓						✓	✓	✓	✓	✓			✓				Premium	
	ResearchKick	✓						✓	✓	✓	✓	✓			✓			✓	Premium	
Graph-Based	Connected Papers	✓		✓				✓											Freemium	214 million
	ScholarGPS	✓			✓	✓	✓	✓											Free	over 200 million
	CiteSpace					✓		✓											Freemium	
	Sci2							✓											Free	
	NLP KG	✓		✓	✓		✓	✓											Free	
	ORKG ASK	✓		✓							✓				✓				Free	76 million
Paper Chat	ChatGPT	✓						✓	✓	✓	✓	✓			✓	✓			Freemium	10 pdf files
	Claude	✓						✓	✓	✓	✓	✓			✓	✓			Freemium	5 pdf files
	Deepseek	✓						✓	✓	✓	✓	✓			✓	✓			Free	
	Research			✓				✓	✓		✓	✓			✓				Freemium	1 pdf file
	NotebookLM							✓	✓		✓	✓			✓			✓	Freemium	50 pdf files
	Enago Read	✓		✓				✓	✓		✓	✓			✓			✓	Freemium	1 pdf file
	DocAnalyzer.AI			✓				✓	✓		✓	✓			✓	✓		✓	Premium	few pdf files
	CoralAI			✓				✓	✓		✓	✓			✓				Freemium	1 pdf file
	ExplainPaper							✓	✓		✓	✓			✓				Freemium	1 pdf file
	ChatPDF	✓		✓				✓	✓		✓	✓			✓				Premium	1 pdf file
Recommender	Arxiv Sanity	✓	✓	✓														✓	Free	
	Scholar Inbox	✓	✓	✓		✓		✓							✓			✓	Free	
	ResearchTrend.ai	✓				✓													Freemium	
	TrendingPapers	✓	✓			✓					✓				✓			✓	Free	
	Bytez	✓				✓			✓	✓		✓	✓		✓	✓			Freemium	
	Notesum.ai	✓	✓	✓							✓				✓			✓	Freemium	
	Research Rabbit	✓		✓			✓												Free	

Other AI-enhanced Literature Search

	Platform	Search	Recommendations	Collections	Citation Analysis	Trending Analysis	Author Profiles	Visualization Tools	Paper Chat	Idea Generation	Paper Writing	Summarization	Paper Review	Datasets	Code Repositories	LLM Integration	Web API	Personalization	Cost	Data Source
Search Engines	Google Scholar	✓	✓	✓	✓		✓											✓	Free	
	Semantic Scholar	✓	✓	✓	✓	✓	✓		✓		✓				✓	✓	✓	✓	Free	214 million
	Baidu Scholar	✓	✓	✓	✓	✓	✓								✓			✓	Freemium	680 million
	BASE	✓		✓												✓			Free	415 million
	Internet Archive Scholar	✓														✓			Free	35 million
	Scilit	✓		✓	✓		✓												Free	172 million
	The Lens	✓		✓			✓									✓			Freemium	284 million
	Science.gov	✓						✓											Free	several million
	Academia.eu	✓		✓			✓												Freemium	55 million
	OpenAlex	✓					✓									✓			Freemium	
	AceMap	✓			✓	✓	✓	✓						✓					Free	260 million
	PubTator3	✓		✓	✓											✓			Free	6 million
Benchm.	Papers with Code	✓												✓	✓				Free	154 thousand
	ScienceAgentBench									✓				✓	✓	✓			Free	
	ORKG Benchmarks					✓		✓						✓					Free	
	Huggingface	✓		✓		✓								✓	✓				Freemium	

Conclusion

- AI is transforming literature search.

Conclusion

- AI is transforming literature search.

- AI tools boost discovery but still require oversight.

Conclusion

- AI is transforming literature search.
- AI tools boost discovery but still require oversight.
- Four AI paradigms jointly redefine research workflows.

Conclusion

- AI is transforming literature search.
- AI tools boost discovery but still require oversight.
- Four AI paradigms jointly redefine research workflows.
- Future directions point to smarter, multimodal systems.

Future Direction

- Multimodal literature search
  - Integrating text and figures/tables
- Event-oriented summarization
  - Extracting and organizing key research events (e.g., discoveries, methods, results) for clearer insights
- Real-time updates & knowledge tracking
  - Continuous integration of new findings
- Integration with scientific knowledge graphs
  - Structured, interconnected research data

Thank you! Any questions?